Moralization in social networks and the emergence of violence during protests.

Marlon Mooijman[1,2], Joe Hoover[1,3,4], Ying Lin[5], Heng Ji[5] & Morteza Dehghani[3,4,6*]

[1]*Both authors contributed equally to this work.*

[2]*Department of Management and Organizations, Kellogg School of Management, Northwestern University.*

[3]*Brain and Creativity Institute, University of Southern California.*

[4]*Psychology Department, University of Southern California.*

[5]*Department of Computer Science, Rensselaer Polytechnic Institute.*

[6]*Department of Computer Science, University of Southern California.*

In recent years, protesters in the US have clashed violently with police and counter-protesters on numerous occasions [1–3]. Despite widespread media attention, little scientific research has been devoted to understanding this rise of violent protest. We propose that this phenomenon can be understood as a function of individual's moralization of a cause and the degree to which they believe others in their social network moralize that cause. Using data from the 2015 Baltimore protests, we show that not only did the degree of moral rhetoric used on social media increase on days with violent protests, but also that the hourly frequency of morally relevant tweets predicted the future counts of arrest during protests, suggesting an association between moralization and protest violence. To better understand the structure of this association, we ran a series of controlled behavioral experiments demonstrating that

people are more likely to endorse violent protest for a given issue when they moralize the issue; however, this effect is moderated by the degree to which people believe others share their values. We discuss how online social networks may contribute to inflations of protest violence.

Protest is widely seen as an important component of democratic societies. It enables constituents to express grievances, communicate directly with the public and representatives, and promote change in accordance with their beliefs. While protests associated with popular platforms often attract large numbers of attendees, they are frequently peaceful events, even when they target controversial issues. Influential theories on social movements suggest that people engage in peaceful protests for a plethora of reasons, including rational deliberations, identification with a political cause, and feelings of relative deprivation[4–6].

However, protests can also quickly erupt into violence. For instance, in the United States, protesters clashed violently with police in Ferguson, Missouri after the killing of Michael Brown[1]. More recently, far-right protesters clashed violently with counter-protesters in Charlottesville, Virginia in response to the proposed removal of a confederate statue[2]. How do we understand, and predict, the acceptability of violent protests and the emergence of violent behavior at protests?

While many factors likely contribute to a protest's risk for violence, we propose that it can be in part understood as a function of two key risk factors: (a) the degree to which people see protest as a moral issue and (b) the rate of perceived moral convergence –the degree to which participants believe that others share these moral attitudes. Further, we suggest that reliable measurements

of these risk factors (moralization and moral convergence) can be obtained from online social network activity; we find not only that violent protest is preceded by an increase in online moral rhetoric, but also that hourly signals of online moral rhetoric predict future hourly arrest counts during violent protests.

We focus on morality because once a protest is sufficiently moralized, it becomes an issue of right and wrong instead of mere personal preference (e.g., mere liking/disliking, mere approval/disapproval, or mere support/non-support for a protest [8–10]). Seeing a protest as a moral issue thus means that people's attitudes about the protest are more absolute and less subject to change[11], with moralization fostering the feeling that something "ought" to be done one way or the other, thereby potentially contributing to the endorsement of protest violence[12]. As not all protests are moralized to the same extent across time, place, and people, variance in moral attitudes can be measured and used to predict violence at protests.

Indeed, our hypotheses are grounded in the observation that protests are often preceded by extensive discussions on Twitter and Facebook about moral topics such as societal unfairness and injustice[13]. Social media platforms, in other words, have become important tools for people to express their moral disapproval with social and political developments such as government corruption, the killing of unarmed citizens by police, and the removal of culturally meaningful symbols and statues[14,15]. Due to the scale of social networks, messages that contain references to moral terms, such as injustice and unfairness, are likely to spread to thousands if not millions of others and reflect the moral sentiments of a given population[16]. Thus, social media discourse materially

encodes signals of moralization and moral outrage[17].

Importantly, such signals are often not what one might refer to as mere rhetoric, as moral sentiments can provide the foundation for violence[18]. For instance, the perceived moral obligation to oppose a political candidate, discipline one's child, or prevent a criminal from reoffending is perceived by some to justify the use of force and violence[19,20]. Suicide bombers kill themselves and others in the name of a divine authority that commands obligation[18,21]. When attitudes towards a social or political issue are seen as a reflection of moral beliefs, people are more willing to use violence to protect their moral beliefs and achieve desired ends[12]. This association between violence and moral beliefs provides the foundation for the first risk factor in our theory: violence at protests is associated with protest-relevant moral rhetoric on online social networks, because this rhetoric reflects moralization, and moralization is a risk factor for violence[12].

However, while moralization is a risk factor for violence, we propose that in isolation it is not always sufficient for violent protest. Rather, the acceptability of protest violence is contingent on both moralization and the belief that others share one's moral attitudes, a phenomenon which we refer to as perceived moral convergence. When people encounter others who share their moral attitudes, those attitudes are validated and reinforced and, as moral beliefs become more intransigent, the likelihood of advocating or enacting violence to achieve desired moral ends (e.g., topple a corrupt government, alter policing practices, stop the removal of a statue, or defend the purity of one's race) may increase[22].

Accordingly, we propose that the risk of violent protest is not simply a function of moraliza-

tion, but also the perception of moral convergence, which we believe can be influenced by social media dynamics. People not only rely on social media to signal their moral sentiments; they also use social media to gauge the moral sentiments held by others. Thus, as morally-relevant messages proliferate, impressions of convergence may become more robust and make people overcome any personally held objections to using violence. Notably, such impressions can be dramatically biased given the tendency for online social networks to function as digital echo chambers[23]. Nonetheless, we suggest that the count of moral rhetoric in online social networks encodes both a signal of moralization and moral convergence.

To test these hypotheses, we collected tweets relevant to the 2015 Baltimore protests. These protests, which were motivated proximally by the eventual death of Freddie Gray, extended across multiple weeks and included both periods of peace and violence, allowing us to examine the links between moral rhetoric and violent protests. Second, across three behavioral experiments, we directly tested whether the interaction between moralization and perceived moral convergence explains support for violent protest. Thus, although the behavioral experiments measure the acceptability of using violence at protests instead of protest behavior, the fine-grained text analysis of the Baltimore protests and the three behavioral experiments taken together aim to provide converging evidence for our hypotheses using both real-life protests and self-reported attitude measures.

To test the hypothesized association between moralization and violent protest, we conducted

an observational study of the relationship between online moral rhetoric and real-world indicators of violent protest during the 2015 Baltimore protests. To operationalize online moral rhetoric, we used 4800 tweets hand-coded for moral content to train a deep neural network (see Study 1: Moral Rhetoric Classification under Methods for more details). We then used this network to predict binary 'moral' or 'not-moral' labels for 18 million tweets posted during the Baltimore protests in cities where a protest responding to the death of Freddie Gray occurred. Using these predicted labels, we investigated whether expressions of online moral rhetoric increase on days with violent protests, relative to days without. Finally, we conducted two fine-grained hour-level analyses investigating the association between counts of online moral rhetoric and counts of arrests in the Baltimore area as reported by the Baltimore police.

To evaluate the association between online moral sentiment and violent protest we conducted three sets of analyses. In the first analysis (Study 1A), we investigated the association between the count of moral tweets and protest violence at the day level. Then, in the second analysis (Study 1B), we evaluated bi-directional Granger causality[24] between the count of moral tweets and the count of arrests made in Baltimore at the hour level. Finally, we estimated a Negative Binomial Bayesian hierarchical time-series model (Study 1B) in order to directly estimate the association between hour-level moral tweet counts and arrest counts.

More specifically, for the first set of analyses, we investigated whether more moral tweets were posted on days with violent protests. To do this, we estimated the daily count of moral tweets (see Study 1A: Design and Analysis for discussion of modeling counts vs. rates) as a negative

binomial function of whether a violent protest occurred or not [25]. To minimize the risk of mistaking variation in the total number of tweets as meaningful variation in the count of moral tweets, we first evaluated the association between the total count of tweets as a negative binomial AR(1) function of violent protest. Results indicated that the total count of tweets on days with violent ($b = 0.08, SE = 0.0007, p < 0.001$) and peaceful ($b = 0.07, SE = 0.0005, p < 0.001$) protests only increased by 8 and 7%, respectively. Notably, these estimates suggest a minimal difference between the volume of tweets on days with violent protests and peaceful protests; further, these effects are small in magnitude, suggesting that (as reflected in our corpus) there was not a large increase in twitter volume on days with protests.

Next, a baseline model (Model 1) of the count of moral tweets was estimated. This model included only the intercept and dispersion parameter. Because the dependent variable is a time-series, residual autocorrelation was both expected and observed in this model, Box-Pierce $\chi^2 = 10.26, p = 0.001$ [26]. Subsequent examination of autocorrelation and partial autocorrelation plots indicated an AR(1) [27] structure and the baseline model was refit with a AR(1) component (Model 2). Examination of the Model 1 residuals indicated that the AR(1) sufficiently accounted for the residual autocorrelation observed in Model 1, Box-Pierce $\chi^2 = 10.26, p = 0.58$.

To evaluate the association between daily moral tweet counts and days with violent protests, we conducted an intervention test[28], which tests the null hypothesis that an 'intervention(s)' (i.e. an event) *does not* effect a dependent time-series variable. This test indicated an effect of day-level violent protest on the daily count of moral tweets, such that the count of tweets on days with

violent protests was substantially different from the count of moral tweets on other days, $\chi^2 = 11.48, p = 0.02$. We then estimated a third model (Model 3) by modifying Model 2 to include two dummy coded factors reflecting whether, for a given day, no protest occurred, a peaceful protest occurred, or a violent protest occurred. Convergent with our hypothesis and the intervention test, the parameters estimated in this model indicate a positive association between the occurrence of a violent protest and the daily count of moral tweets, $b = 0.63, SE = 0.2, 95\%CI = [0.24, 1.02]$. That is, on days with violent protests, the log expected counts of moral tweets is expected to increase by 0.63. In terms of incidence ratios, this means that the count of moral tweets on days with violent protests is 1.88 times that of days with no protests, holding the other variables in the model constant. No such association was observed for peaceful protest days, $b = 0.11, SE = 0.14, 95\%CI = [-0.17, 0.39]$ (Figure 1).

Further, to investigate whether this effect might be merely driven by the total volume of daily tweets, we also modified Model 3 to include an offset equal to $log(\frac{DailyTotalTweets}{1000})$ (Model 4). Due to the offset, Model 4 models the rate of moral tweets per 1000 tweets, rather than the *count* of moral tweets. Notably, there are no substantive differences between the estimates obtained from Models 3 and 4. In Model 4, the estimated log-odds effects of violent protests and peaceful protests are ($b = 0.57, SE = 0.11, Z = 4.94, p < 0.001$) and ($b = 0.03, SE = 0.08, Z = 0.41, p = 0.68$).

This analysis provides evidence for the operational hypothesis that days with violent protests have higher counts of moral tweets. That is, if a moral protest occurs on a given day, this model indicates that we should expect the count of moral tweets to also increase on that day. Further, we

Figure 1: Expected and observed daily moral tweet counts by protest type. The shaded band around the expected trend line indicates the 95% HPD interval.

observed only a weak relationship between the total volume of daily tweets and protest and models of both the daily count and rate of moral tweets yielded consistent results. Thus, the observed effects do not appear to be driven merely by variation in the total volume of tweets.

However, while these results support our hypotheses, they do not constitute direct confirmatory evidence. For example, the count of moral tweets might increase on violent protest days because people tweet about the violent protests after they occur. Thus, a better question is, "does moral rhetoric predict violence at protests?" In the next study, we address this question by conducting a more fine-grained time-series analysis. Specifically, we investigate whether the count of moral tweets predicts the count of arrests during protests.

The previous study (Study 1A) provided evidence for the hypothesis that the count of moral tweets increases on days with violent protests. Here (Study 1B) we tested the hypothesis that the hourly count of moral tweets predicts protest violence. However, because hourly estimates of violence during the Baltimore protests do not exist, it was not possible to rely on direct measurements of violence. To overcome this issue, we used hourly arrest counts in the Baltimore area as an indicator of protest violence. While arrest counts are an imperfect indicator of protest violence (e.g. arrests can happen without violence and vice versa; arrests happen *after* violence and this gap likely varies), they do provide a dynamic, albeit imperfect, indication of protest violence and thus

offer a unique and valuable opportunity to conduct a fine-grained test of our hypothesis.

To evaluate the relationship between hourly arrest counts and hourly moral tweets, we relied on two complementary modeling frameworks. First, we used the Toda and Yamamoto procedure[33] to conduct tests of Granger causality[24]. An independent variable is said to 'Granger cause' a dependent variable when previous values of the IV predict future values of the DV above and beyond predictions based on past values of the DV alone. The Toda and Yamamoto procedure facilitates testing for Granger causality between two non-stationary time-series variables, which present challenges for conventional tests of Granger causality (See Study 1B: Design and analysis for details on the Toda and Yamamoto process).

The results of this analysis were consistent with the hypothesis of Granger causality, $\chi^2(2) = 17.5, p = 0.0002$. That is, the standardized log count of moral tweets Granger causes the count of arrests. Further, tests for Granger causality in the opposite direction (i.e. flowing from arrest counts to moral tweet counts) were also rejected, $\chi^2(2) = 6.4, p = 0.04$. Accordingly, these analyses indicate a bi-directional Granger causal relationship, such that the count of moral tweets predicts the future count of arrests and the count of arrests predicts the future count of moral tweets.

While the results of our Granger causality analysis were consistent with our hypothesis, a potential shortcoming is the possibility for temporal gaps between protest violence and resulting arrests. Further, Granger causality analysis does not aim to estimate the magnitude of effects, which in our case are of particular interest. Accordingly, Negative-Binomial $AR(2)(1)[24]$ models were used to estimate the relationship between hourly arrest counts during a given hour and the

average count of moral tweets during the previous four hours (See Study 1B: Design and analysis for more details on the identification of this AR structure and model selection). By focusing on the average count of moral tweets over a four-hour window, we aimed to relax any rigid assumptions about the temporal association between moral tweets and arrests.

The first model was estimated using the $tscount$ package[34] and the effect of standardized log count of moral tweets was treated as fixed. In this model, the expected effect of the moral tweets variable ($b = 0.22, SE = 0.09, 95\%CI = [0.04, 0.39]$) indicated that the count of moral tweets across a four hour window predicts the count of arrests in the next hour. The second version of this model, which was estimated using Bayesian estimation and which allowed the intercept and moral tweets effect to vary across days, revealed substantial between-days variation for the intercept ($SD = 0.20$) and the effect of the moral tweets variable ($SD = 0.16$). However, the estimate of the fixed effect of the moral tweets variable was comparable to the previous model ($b = 0.24, SD = 0.10, 95\%HPDI = [0.07, 0.44]$). Thus, even after accounting for arrest counts during previous hours, variable baseline arrest counts across days, and potential variation in the effect of moral tweets on arrest counts, these models suggest that for a one unit increase of the average standardized log count of moral tweets over a four hour time span, the log count of arrests is expected to increase by 0.24. In terms of incidence ratios, this means that the count of arrests increases by 1.27 for every one unit increase in the moral tweets variable (see Figure 2B for observed and predicted hourly arrest counts).

Conjointly, these analyses indicate that the count of moral tweets predicts the count of ar-

rests. We find evidence for this both via tests of Granger causality and direct estimates of the association between lagged moral tweets and number of arrests. Specifically, the Granger causality analysis found that moral tweet counts predicted future arrest counts above and beyond current arrest counts. Further, we found evidence for our directional hypothesis; as the count of moral tweets increases, the expected future count of arrests also increases. Thus, even after accounting for arrest counts during previous hours, variable baseline arrest counts across days, and potential variation in the effect of moral tweets on arrest counts, our model indicates a relationship between the count of moral tweets and the future count of arrests. In other words, this model suggests that observing expressions of moral sentiment in social media can help predict when future protests will take on the sort of characteristics that lead to higher counts of arrests.

However, due to the high noise-to-signal ratio in social media data, the small time-span covered by the Baltimore Protest data, and the considerable variance in the models prediction intervals, these results should be interpreted as suggestive rather than constituting conclusive evidence for a relationship between moral sentiment and violent protests. Also, while these results correspond to our general hypothesis, they do not address the question of what, exactly, higher counts of moral sentiment indicate. As discussed above, an increase in moral sentiment could simply reflect an increase in moralization; however, given that people rely on social media to track popular opinions, it may also be the case that an increase in moral sentiment also indexes perceived convergence. Accordingly, we followed this study with three experimental studies, which allowed us to conduct more precise hypothesis tests and to disentangle the potential effects of moralization and moral convergence.

Figure 2: A: Z-Scored time-series of smoothed hourly counts of arrests and moral tweets. B: Observed and predicted arrest counts. The gray band indicates the 95% HPDI predictive interval.

In Study 2, we tested the degree to which the moralization of a protest predicted the acceptability of using violence at this protest. Specifically, we used the violent protests between the far-right and counter-protesters in Charlottesville, Virginia in August 2017 as an example. We tested to what extent protesting the far-right was seen as a moral issue and to what extent this moralization predicted the perceived acceptability of using violence against the far-right.

We introduced participants ($N = 275$) with a short description of the Unite the Right rally in Charlottesville, Virginia, United States, where far-right protesters and counter-protesters clashed. Specifically, participants read the following: "The Unite the Right rally (also known as the Charlottesville rally) was a far-right rally in Charlottesville, Virginia, United States, from August 11-12, 2017. The rally occurred amidst the backdrop of controversy generated by the removal of several Confederate monuments". Participants were informed that we were interested in their opinion about the counter-protesters that protested the far-right protesters. Participants indicated on a four item scale (1 = disagree completely, 5 = agree completely; $M = 3.45, SD = 1.25$) to what extent they thought that protesting the far right was a moral issue (i.e., whether protesting the far right was: a reflection of their moral beliefs and convictions, connected to their beliefs about moral right and wrong, based on moral principle, and a moral stance[12], Cronbach's α = .94). We also measured participants' political orientation on a scale from 1 (extremely liberal) to 9 (extremely conservative; $M = 4.24, SD = 2.16$). This single-item measure of ideology is frequently used

and exhibits strong predictive validity (e.g.,[35]).

Participants then indicated (1 = disagree completely, 7 = agree completely; $M = 2.74, SD = 1.35$) to what extent they agreed with the following six statements: it is acceptable to use violence against far-right protesters, the use of violence against far-right protesters is justified, violence against far-right protesters is acceptable if its means fewer future protests from the far-right, using force during a protest is wrong even if its leads to positive change, using force during a protest against the far-right is immoral even if its leads to positive change, and using violence against the far-right is unacceptable (the last three items were reverse-coded; α = .83).

A regression analysis in which violence was regressed on moralization showed that, as expected, the moralization scale was positively associated with the acceptability of using violence at the protest ($b = 0.22, SE = 0.06, t[274] = 3.44, p = .001, 95\%CI = [0.09, 0.35]$), even when controlling for the political orientation of participants ($b = 0.21, SE = 0.07, t[274] = 3.10, p = .002, 95\%CI = [0.08, 0.35]$). Political orientation did not significantly correlate with the acceptability of using violence at this protest ($r = -.09, p = .15$) and did not predict the acceptability of violence when added to the regression model with the moralization scale ($b = -0.01, SE = 0.04, t[275] = -0.22, p = .82, 95\%CI = [-0.09, 0.07]$). These findings suggest that moralization is indeed associated with the increased acceptability of violent protests.

Study 3 aimed to replicate Study 2 while also disentangling the potential effects of moralization and moral convergence on violence acceptability.

Participants ($N = 201$) were confronted with the same excerpt and moralization questionnaire as in Study 2 (α = .91; $M = 2.74, SD = 0.92$). In addition, in the high-convergence [low-convergence] condition, participants were told that based on their responses, "the majority of [few] people in the United States share your particular moral values. Other people in the United States think about this protest in a similar [different] manner compared to you". Similar to Study 2, participants then indicated on six items (1 = disagree completely, 7 = agree completely; $M = 3.12, SD = 1.59$) to what extent they considered using violence against far-right protesters acceptable, α = .89).

Regression analysis was used to test the interactive effects of moralization and moral convergence on the perceived acceptability of violence. For the first step, moralization and moral convergence were added. For the second step, the interaction between moralization and convergence was added. Results demonstrated that moralization was positively associated with the acceptability of violence ($b = 0.29, SE = 0.12, t[201] = 2.41, p = .017, 95\%CI = [0.05, 0.53]$) and that moral convergence was (overall) unrelated to the acceptability of violence ($b = 0.05, SE = 0.22, t[201] = 0.21, p = .84, 95\%CI = [-0.39, 0.49]$). Crucially, we observed a significant interaction effect between moralization and moral convergence ($b = 0.48, SE = 0.24, t[201] = 1.98, p = .049, 95\%CI = [0.01, 0.95]$). Moralization predicted the acceptability of violence at protests when moral convergence with others was high ($b = 0.52, SE = 0.18, t[99] = 2.97, p = .004, 95\%CI = [0.17, 0.87]$) but not when moral convergence with others was low ($b = 0.05, SE = 0.16, t[101] = 0.27, p = .79, 95\%CI = [-0.28, 0.37]$).

Similar to Study 2, findings from Study 3 indicate that participants found violence more acceptable when they moralized a protest. In addition, Study 3 shows this effect to be moderated by the degree to which people believed others shared their moralized attitudes. Moralization predicted violence only when participants perceived that they shared their moralized attitudes with others.

In Study 4 we aimed to directly replicate the findings from Study 3 while also measuring attitude certainty as a possible explanation for why moralization primarily predicts the acceptability of violence under conditions of moral convergence.

Participants ($N = 289$) were confronted with the same excerpt and moralization questionnaire as in Studies 2 and 3 (α = .95; $M = 3.75, SD = 1.19$), and the same convergence manipulation as in Study 3. Similar to the previous experiments, participants then indicated on six items (1 = disagree completely, 7 = agree completely) to what extent they considered using violence against far-right protesters acceptable (α = .79;$M = 3.49, SD = 1.33$). Participants also indicated to what extent they were certain about their attitude towards the protest (1 = extremely uncertain, 7 = extremely certain; $M = 5.60, SD = 1.60$).

We conducted similar analyses as in Study 3. Results demonstrated that moralization was positively associated with attitude certainty ($b = 0.23, SE = 0.08, t[289] = 2.93, p = .004, 95\%CI = [0.07, 0.38]$) and violence acceptability ($b = 0.28, SE = 0.06, t[289] = 4.41, p < .001, 95\%CI = [0.15, 0.40]$). Moral convergence was (overall) positively related to attitude certainty ($b = 0.49, SE = 0.18, t[289] = 2.66, p = .008, 95\%CI = [0.13, 0.85]$) and violence acceptability ($b = 0.32, SE = 0.15, t[289] = 2.12, p = .035, 95\%CI = [0.02, 0.62]$).

Crucially, we observed a significant interaction effect between moralization and moral convergence for attitude certainty ($b = 0.33, SE = 0.15, t[289] = 2.14, p = .034, 95\%CI = [0.03, 0.63]$) and violence acceptability ($b = 0.42, SE = 0.12, t[289] = 3.40, p = .001, 95\%CI = [0.18, 0.67]$; see Figure 3). Moralization predicted attitude certainty and violence acceptability when moral convergence with others was high ($b = 0.39, SE = 0.09, t[144] = 4.24, p < .001, 95\%CI = [0.21, 0.57]$; $b = 0.49, SE = 0.09, t[144] = 5.41, p =< .001, 95\%CI = [0.31, 0.67]$, respectively) but not when moral convergence with others was low ($b = 0.06, SE = 0.12, t[144] = 0.49, p = .62, 95\%CI = [-0.18, 0.30]$; $b = 0.07, SE = 0.09, t[144] = 0.76, p = .45, 95\%CI = [-0.10, 0.23]$, respectively).

In addition, a 5,000 resamples bootstrapping analysis[36] demonstrated that attitude certainty ($95\%CI = [0.01, 0.13]$) mediated the interaction effect between moralization and moral convergence on violence acceptability. Specifically, the indirect effect of attitude certainty was significant when moral convergence was high ($95\%CI = [0.03, 0.12]$) but not when moral convergence was low ($95\%CI = [-0.03, 0.05]$). Indeed, attitude certainty overall predicted the acceptability of violence ($b = 0.22, SE = 0.05, t[289] = 4.60, p < .001, 95\%CI = [0.13, 0.31]$) and this was the case to a greater degree when moral convergence was high ($b = 0.31, SE = 0.08, t[144] = 3.84, p < .001, 95\%CI = [0.15, 0.47]$) compared to low ($b = 0.14, SE = 0.06, t[144] = 2.48, p = .014, 95\%CI = [0.03, 0.25]$).

Studies 2, 3, and 4 confirm that the moralization of a protest can increase the acceptability of using violence at this protest. In addition, this only occurred when participants perceived

Figure 3: Acceptability of using violence during a protest as a function of moralization and moral convergence for Study 4. Error bars represent standard errors.

to share their moralized attitudes with others and increased their attitude certainty. Overall, the findings from these three behavioral experiments are consistent with the findings from the Baltimore protests such that the moralization of protests and the moral convergence of these moralized attitudes drives the acceptability of using violence at protests.

Taken together, the findings from 18 million tweets posted during the 2015 Baltimore protests and three behavioral experiments suggest that violence at protests can be understood as a function of two key risk factors: the degree to which people see protest as a moral issue and the degree of perceived moral convergence. A rise in violence at protests may thus reflect the increasing moralization and polarization of political issues in online echo chambers. The risk of violent protest, in other words, may not be simply a function of moralization, but also the perception that others agree with one's moral position, which can be strongly influenced by social media dynamics. Indeed, previous research has shown that people do not just rely on social media to signal their moral sentiment but they also use social media to gauge the moral sentiments held by others[23,37].

As morally-relevant messages proliferate in social networks, impressions of convergence may become more robust, as suggested by the findings reported in the current manuscript. This may increase the degree to which people overcome their objections to using violence aimed at

perceived opponents[20]. Notably, impressions of moral convergence on online social networks can be dramatically biased given the tendency for online social networks to function as digital echo chambers[23]. It is estimated that seven out of ten Americans are connected to some online social network[38] and that political polarization has been steadily increasing in recent decades[39], implying that online social networks are currently playing a significant role in shaping, and perhaps causing, our attitudes surrounding the use of violence aimed at ideological opponents at protests.

These findings come at a time when some polls suggest that a minority of US college students consider it acceptable to use violence at protests[3]. In the last few years, protesters in the US have clashed violently with police and counter-protesters on numerous occasions, whether this was due to the killing of young men by police, protests by the far-right, or the invitation of controversial speakers on college campuses. The findings reported in the current manuscript shed some light unto these social development while also providing suggestions for counteracting the increasing acceptability of violence. Increased support for using violence at protests occurred only when people perceived to share moralized attitudes with others but not when people did not perceive to share moralized attitudes with others. This implies that decreasing the moralization of attitudes and diluting the perception that others agree with one's moral position may attenuate the rise of the acceptability of violence. Although people may try to "sort" themselves into networks with high levels of moral convergence, our findings imply that ways of combating this may be effective at decreasing the acceptability of using violence at protests. In particular, convergence is most relevant in relation to our direct social circle as this is where we derive most of our sense of identity and meaning from[40]. As such, future research could investigate whether convergence

impacts relevant outcomes for epistemic (if my peers agree with me, I must be right) or social (if my peers disagree with me, then I can get in trouble for acting on my beliefs) reasons.

However, while we propose that the combination of moral outrage and perceived moral convergence are prerequisites for violent protest, they are by no means sufficient conditions for violent protest. In addition to these factors, whether a protest will become violent or not might also depend on a range of other contextual factors, such as the propensity for violence among the protesting population (i.e., base-count levels of violent inclinations), likelihood of instigation from involved non-protesting agents, and the specific nature of the issues being protested (e.g., a march explicitly aimed at being peaceful). Nonetheless, for any given configuration of contextual factors, the current work suggests that the risk of violence increases as a function of moralization and perceived convergence of moralization. Indeed, our work extends previous research on (sacred) values and violence by suggesting that strong moral convictions can increase the acceptability of violence[21,41] and that (a) this happens primarily when convergence is high, and (b) the link between moralization and violence can be tracked, and is shaped, by online social networks.

As a consequence, our findings have far-reaching practical implications, as a key decision-making problem for government officials is to understand which protests will turn violent and how resources should be allocated to prevent protests from spiraling out of control[42]. Although more research should be done to replicate and extend the findings reported in the current manuscript, we believe that our findings suggest that online social networks and moral psychology can be of some help: Given the increasing importance of social networks in our daily lives, the moral language

used on online social networks can be directly linked to violent protests, implying that online social networks can also be used by policy makers to track and predict the emergence of violence at protests.

Our findings also offer a warning about the potential effects of perceived versus actual moral and political homogeneity. Perceived moral and political homogeneity is likely to be higher than actual homogeneity in attitudes, as social media often acts as an ideological echo chamber: people tend to to use their social networks to be in contact with similarly-minded others. It might, then, be worthwhile to investigate the perceived and actual homogeneity of attitudes and confront people with the potential discrepancy between their perceived and actual homogeneity of attitudes. This may prevent a potential slide to violence towards ideologically dissimilar others. Lastly, aspirations for morally homogeneous societies have existed throughout human history (e.g., [43,44]). Our findings hint at potential dangerous consequences of such utopian uniformity.

Methods

Data availability The data files that support the findings of the studies are available at `https://osf.io/wqzmj/files/`. For Study 1, due to restrictions set by Twitter, the ID's of the tweets (and not the tweet texts) have been made available. The publicly available Twitter API can be used to retrieve the original texts of the tweets using the tweet IDs. The (human) annotated tweets have been uploaded for each individual annotator, and for the intersection of the annotators.

Code availability All code used in this paper are available at `https://osf.io/wqzmj/files/`.

Study 1: Data Collection. Approximately 19 million tweets posted during the Baltimore protests (04/12/2015 to 05/08/2015) were purchased from Gnip.com. These tweets were filtered based on geolocation information and constrained to cities where protests related to the death of Freddie Gray occurred. Compared to focusing on a specific list of hashtags, filtering by geo-location allows us to focus more holistically on the nation-wide social stream rather than on some small portion associated with experimenter-determined hashtags.

Study 1: Moral Rhetoric Classification. To evaluate the association between online moral rhetoric and violent protest, it was first necessary to measure the moral sentiment of the tweets in the Baltimore corpus. We accomplished this using a Long Short-Term Memory neural network trained on a set of 4800 tweets labeled by expert annotators[46]. We first developed a coding manual[47] and trained three human annotators to code the tweets for moral content based on the Moral Foundations Theory[48–50]. Specifically, each coder was asked to annotate each tweet depending on whether it was related to any of the five Moral Foundations or not. Agreement was measured using Prevalence and Bias adjusted Kappa (PABAK)[51,52], an extension of Cohen's Kappa robust to unbalanced data sets. The calculated agreement was high for all dimensions (for the moral dimensions averaged across coder pairs, $M = 0.723, SD = 0.168$). The agreement for moral/non-moral categories was 0.636.

The annotated tweets were then used to train a deep neural network-based model to auto-

matically predict moral values involved in Twitter posts[46]. This model consists of three layers, an embedding (lookup) layer, a recurrent neural network (RNN) with long short-term memory (LSTM)[53] and an output layer. The first layer converts words in an input tweet to a sequence of pre-trained word embeddings, which are low-dimensional dense vectors that represent semantic meanings of words. After that, the LSTM layer processes these embeddings in succession and outputs a fixed-sized vector which encodes critical information for moral value prediction. Compared to vanilla RNNs, an LSTM is capable of storing inputs over long sequences to model long-term dependencies. Other dense features, such as a vector representing percentage of words that match each category in the Moral Foundation Dictionary[48], are concatenated with the LSTM feature vector. On top of the model, a fully connected layer with Softmax activation function is added to perform binary classification. It takes as input the concatenated feature vector and predicts whether a tweet reflects a moral concern or not. We trained a separate model for each moral foundation and combined all results. Previous research has found that the moral content labels predicted by this model are generally not distinguishable from labels generatedd by humans[46]. Our model achieved a cross-validated accuracy of 89.01% (F_1 = 87.96) compared to 75.15% (F_1 = 72.17) of Moral Foundations Dictionary[48] on the tweets used for training.

Study 1A: Data. Using the LSTM Tweet Categorization Model (see above), we generated binary labels indicating the presence of each Moral Foundation domain. We then calculated the hour-level mean moral tweet count across Foundations, which we use our as hour-level dependent variable. Day-level moral tweet count was calculated by summing the hourly-averages within days. Then, for each day in the range of dates included in the twitter corpus, media coverage of the timeline of

the Baltimore protests was used to label each day for whether it had no protest (n = 12), a peaceful protest (n = 11), or a violent protest (n = 4).

Study 1A: Design and Analysis. Because the dependent variable (daily count of moral tweets) is a count variable and overdispersed ($M = 8188.45, s^2 = 23,065,122$), we used negative binomial regression. In total, three models were estimated: a null, intercept-only model (Model 1); a model with a sufficient serial autocorrelation structure (Model 2); and a model with both an autocorrelation component and dummy variables indicating protest type (Model 3). Unless otherwise stated, all models in Studies 1A and 1B were estimated in the R statistical computing environment version 3.3.1[54] using the package $tscount$ version 1.4.0[34].

Often, when modeling count variables it is necessary to account for variation in total population. That is, what is of interest is not the *count* of a particular outcome, but its rate. Within the context of general linear models, this is typically achieved by including a so-called 'off-set' [60] term, which accounts for variations in total population (conventionally referred to as exposure) by fixing the off-set parameter to 1 [60]. However, in the context of the current work, it is not clear that accounting for variations in exposure (e.g. the total number of tweets) is more valid than directly modeling the count of moral tweets, because the number of tweets is variable and impacted by factors irrelevant to our analysis. For example, imagine that the total count of tweets decreases from day 1 to day 2, but that the count of moral tweets stays constant. In this case, the rate of moral tweets *increases*; however, it is not at all clear that this increase is the kind of increase we are interested in. For example, if the *decrease* in *total tweets* was driven by factors irrelevant to our analyses (e.g. if it is just noise in the Twitter feed), our view is that this is not relevant variation in

the rate of moral tweets *not*.

Accordingly, rather than modeling the rate of moral tweets, we directly model the count of moral tweets. Further, to minimize the risk of mistaking an effect of exposure (i.e. total tweets) for an effect of moral tweets, we estimate the association between the total count of tweets and protest. If an effect of moral tweets is merely masking an effect of total tweets, the estimated effect of the former should be bounded by the latter. Thus, by estimating the association between total tweets and protest, we can establish a minimum effect threshold. Finally, to further minimize the risk misinterpreting a mere effect of total tweets, we compare models of both moral tweet counts and rates.

To test our hypothesis, we first used Model 1 to conduct an intervention test for count time-series data[28], which tests the null hypothesis that a specified intervention has no effect on the data generation process (DGP). In our case, the alternative hypothesis assumes that days with violent protests statistically intervene on the count of moral tweets. Functionally, this test treats intervention time-points (i.e. violent protest days) as potential outliers and estimates the likelihood that they were generated from the baseline DGP. If the intervention time-points appear to have come from a different DGP, this counts as evidence for an intervention effect. We follow this intervention test with Models 3 and 4, which directly estimate the effects of violent protest days on the daily count and rate of moral tweets.

Study 1B: Data. Hourly counts of moral tweets were calculated using the labeled Twitter data described in study 1A. Hourly arrest counts for the Baltimore area were obtained from the Open

Baltimore data portal at `https://data.baltimorecity.gov/`.

Study 1B: Design and Analysis. Both the hourly count of moral tweets and the hourly count of arrests are time-series (see Figure 2A for arrest and moral tweet count time series). Thus, whether the former predicts the latter cannot be determined with conventional regression models because shared temporal trends (such as an seasonal pattern at the hour level) can induce spurious correlations between otherwise independent time-series[27]. One widely-used method for overcoming this problem is the Granger causality test[24]. This test assumes that a dependent variable Y is at least partially a function of some set of endogenous factors and it tests whether an independent variable X provides additional predictive information. In practice, Granger causality is tested by regressing Y at $t + 1$ both on lagged values of Y and X and evaluating whether the lagged X components improve the model. If lagged X values do not predict Y at $t + 1$ above and beyond the degree to which lagged Y values predict Y at $t + 1$, then Granger non-causality is assumed. However, if the lagged X values contribute additional predictive information to the model, then Granger causality is assumed.

To evaluate the relationship between the hourly counts of moral tweets and arrests, we first test for Granger causal relationship between moral tweets and arrests on days with peaceful or violent protests using the Toda and Yamamoto[33] procedure. Because it may also be the case that the count of arrests Granger causes the count of moral tweets, we also test the reverse model. That is, while it may be that the count of moral tweets is an indicator of protest violence, it may also be the case that protest violence granger causes the count of moral tweets.

A Vuong model comparison test[29] indicated that hourly arrest count was better fit by the null Negative Binomial model, compared to a Poisson model. Per the Toda and Yamamoto procedure, the maximum order of integration p across tweet counts and arrest counts ($p = 1$) was identified across both moral tweet counts and arrest counts using augmented Dickey-Fuller[30] and Kwiatowski-Phillips-Schmidt-Shin[31] tests for stationarity. Autocorrelation and partial autocorrelation plots and Box-Pierce[32] tests were then used to identify the maximum number of lags m ($m = 2$). Examination of the residual correlation plots also revealed seasonal correlation at the hourly level. Finally, per the Toda and Yamamoto procedure, a final model with $m+p = 2+1 = 3$ lags for both number of arrests and number of moral tweets was estimated. This model also included a 24 hour arrest count lag to account for hourly seasonality. Finally, the Wald χ^2 test of the null hypothesis of Granger non-causality, that the slopes of the m moral tweet lags are not distinguishable from zero, was rejected, $\chi^2(2) = 17.5, p = 0.0002$. To determine whether the count of arrests Granger causes the standardized log count of moral tweets, the transformed moral tweets variable was entered in a linear regression with the same predictors as the previous model. The Wald χ^2 test of Granger non-causality was also rejected for this model, $\chi^2(2) = 6.4, p = 0.04$.

Following these tests of Granger causality, we then estimate a dynamic time-series model in order to estimate the degree to which the lagged count of moral tweets is associated with the count of arrests. Because we do not have a hypothesis about the specific temporal association between moral tweets and arrests, to estimate this association we average the hourly counts of moral tweets over a four hour window and use this four-hour rolling mean to predict the count of arrests at $t+1$. That is, we estimate the count of arrests at $t+1$ as a function of the average count of moral tweets

across $t-3$, $t-2$, $t-1$, and t. Lastly, we estimate two versions of this model. In the first version (model 1), the effect of moral tweets on arrest count is modeled as a fixed effect. In the second and final version of the model (model 2), we relax this assumption, as well as the assumption of fixed intercepts across days.

Thus, in this final model, we include all of the terms included in the former model, but also allow the intercept and the slope of the moral tweets variable to vary across days. In other terms, we estimate a hierarchical model with so-called random intercepts and slopes. By relaxing the fixed slope assumption, this model better reflects the uncertainty of the slope estimates[55]. This model is estimated using Bayesian estimation via the $rstanarm$ R package version 2.15.3 [56] with weakly informative priors. In all of these models, hourly moral tweet counts ($M = 406.5, SD = 418.92$) was log-transformed and standardized so that its scale was comparable to that of hourly arrest counts ($M = 3.10, SD = 3.57$). Further, as in Study 1A, hourly arrest count is a count variable and cannot be expected to be normally distributed. To allow for the possibility of overdispersion, arrest counts were modeled as Negative Binomial random variables.

Evaluation of the model 1 parameter estimates indicated a positive effect ($b = 0.22, SE = 0.09, 95\% CI = [0.04, 0.39]$) of moral tweet count on arrest counts, such that the count of moral tweets across a four hour window predicts the count of arrests in the next hour. Model 2 marginal estimates of this effect ($b = 0.24, SD = 0.10, 95\% HPDI = [0.07, 0.44]$) were consistent with the model 1 estimates. This consistency was maintained after accounting for substantial variation between day-level intercepts ($SD = 0.20$) and in the effect of moral tweet counts across days ($SD = 0.16$).

Ethical Procedures. All behavioral experiments were approved by the USC IRB panel (UP-17-00375 & UP-CG-16-00006 & UP-16-00682). Prior to participating in the experiments, all subjects were provided an information sheet, approved by the IRB, explaining the studies.

Study 2: Participants. Two hundred and seventy-five participants (169 women; M_{age} = 33.73 years, SD_{age} = 10.61) were successfully recruited from Mechanical Turk (MTurk)[57]. A power analysis indicated that our sample size provided 99% power to detect a medium effect size of 0.25. All power analyses were conducted using G*Power 6[58].

Study 3: Participants. Two hundred and five participants (131 men; M_{age} = 35.04 years, SD_{age} = 11.61) were successfully recruited from Mechanical Turk and randomly assigned to two conditions: high versus low moral convergence. A power analysis indicated that our sample size provided 90% power to detect a medium effect size of 0.25.

Study 4. Participants Two hundred and eighty-nine participants (162 men; M_{age} = 33.78 years, SD_{age} = 9.52) were successfully recruited from Mechanical Turk and randomly assigned to two conditions: high versus low moral convergence. A power analysis indicated that our sample size provided 99% power to detect a medium effect size of 0.25.

1. Davey, M. & Bosman, J. Protests flare after ferguson police officer is not indicted. *The New York Times* (2014). URL https://www.nytimes.com/2014/11/25/us/ferguson-darren-wilson-shooting-michael-brown-grand-jury.html.

2. Heim, J. Recounting a day of rage, hate, violence and death. *The Washington Post* (2017). URL https://www.washingtonpost.com/graphics/2017/local/charlottesville-timeline/?utm_term=.b960304937fe.

3. Rampell, C. Protests flare after ferguson police officer is not indicted. *The Washington Post* (2017). URL https://www.washingtonpost.com/opinions/a-chilling-study-shows-how-hostile-college-students-are-toward-free-speech/2017/09/18/cbb1a234-9ca8-11e7-9083-fbfddf6804c2_story.html.

4. Walker, I. & Pettigrew, T. F. Relative deprivation theory: An overview and conceptual critique. *British Journal of Social Psychology* **23**, 301–310 (1984).

5. Olson, M. *Logic of Collective Action: Public Goods and the Theory of Groups* (Harvard University Press, 1965).

6. Van Stekelenburg, J. & Klandermans, B. The social psychology of protest. *Current Sociology* **61**, 886–905 (2013).

7. McLaughlin, J. & Schmidt, R. National undergraduate study. *McLaughlin & Associates* (2017). URL http://mclaughlinonline.com/2017/10/16/the-william-f-buckley-jr-program-at-yale-survey-30-of-students-believe-

that-physical-violence-can-be-justified-to-prevent-someone-from-using-hate-speech/.

8. Mooijman, M. *et al.* Resisting temptation for the good of the group: Binding moral values and the moralization of self-control. *Journal of Personality and Social Psychology* (2017).

9. Rozin, P. The process of moralization. *Psychological Science* **10**, 218–221 (1999).

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

11. Skitka, L. J., Bauman, C. W. & Sargis, E. G. Moral conviction: Another contributor to attitude strength or something more? *Journal of personality and social psychology* **88**, 895 (2005).

12. Skitka, L. J. & Morgan, G. S. The social and political implications of moral conviction. *Political Psychology* **35**, 95–110 (2014).

13. Manjoo, F. The alt-majority: How social networks empowered mass protests against trump. *The New York Times* (2017). URL https://www.nytimes.com/2017/01/30/technology/donald-trump-social-networks-protests.html?mcubz=0.

14. Steinert-Threlkeld, Z. C. Spontaneous collective action: peripheral mobilization during the arab spring. *American Political Science Review* **111**, 379–403 (2017).

15. Zeitzoff, T. Anger, legacies of violence, and group conflict: An experiment in post-riot acre, israel. *Conflict Management and Peace Science* 0738894216647901 (2016).

16. Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A. & Van Bavel, J. J. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences* **114**, 7313–7318 (2017).

17. Crockett, M. Moral outrage in the digital age. *Nature Human Behaviour* (2017).

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

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

20. Zaal, M. P., Laar, C. V., Ståhl, T., Ellemers, N. & Derks, B. By any means necessary: The effects of regulatory focus and moral conviction on hostile and benevolent forms of collective action. *British Journal of Social Psychology* **50**, 670–689 (2011).

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

22. Boothby, E. J., Clark, M. S. & Bargh, J. A. Shared experiences are amplified. *Psychological science* **25**, 2209–2216 (2014).

23. Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A. & Bonneau, R. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science* **26**, 1531–1542 (2015).

24. Granger, C. W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* 424–438 (1969).

25. Cameron, A. C. & Trivedi, P. K. *Regression analysis of count data*, vol. 53 (Cambridge university press, 2013).

26. McLeod, A. I. & Li, W. K. Diagnostic checking arma time series models using squared-residual autocorrelations. *Journal of Time Series Analysis* **4**, 269–273 (1983).

27. Chatfield, C. *The analysis of time series: an introduction* (CRC press, 2016).

28. Fokianos, K. & Fried, R. Interventions in ingarch processes. *Journal of Time Series Analysis* **31**, 210–225 (2010).

29. Vuong, Q. H. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society* 307–333 (1989).

30. Banerjee, A., Dolado, J. J., Galbraith, J. W., Hendry, D. *et al.* Co-integration, error correction, and the econometric analysis of non-stationary data. *OUP Catalogue* (1993).

31. Kwiatkowski, D., Phillips, P. C., Schmidt, P. & Shin, Y. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics* **54**, 159–178 (1992).

32. Box, G. E. & Pierce, D. A. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association* **65**, 1509–1526 (1970).

33. Toda, H. Y. & Yamamoto, T. Statistical inference in vector autoregressions with possibly integrated processes. *Journal of econometrics* **66**, 225–250 (1995).

34. Liboschik, T., Fokianos, K. & Fried, R. *tscount: An R package for analysis of count time series following generalized linear models* (Universitätsbibliothek Dortmund, 2015).

35. Mooijman, M. & Stern, C. When perspective taking creates a motivational threat: The case of conservatism, same-sex sexual behavior, and anti-gay attitudes. *Personality and Social Psychology Bulletin* **42**, 738–754 (2016).

36. Hayes, A. F., Preacher, K. J. & Myers, T. A. Mediation and the estimation of indirect effects in political communication research. *Sourcebook for political communication research: Methods, measures, and analytical techniques* **23**, 434–465 (2011).

37. Dehghani, M. *et al.* Purity homophily in social networks. *Journal of Experimental Psychology: General* **145**, 366 (2016).

38. Pew Research Center. Social media fact sheet. *Pew Research Center* (2017). URL `http://www.pewinternet.org/fact-sheet/social-media/`.

39. Doherty, C. 7 things to know about polarization in america. *Pew Research Center* (2014). URL `http://www.pewresearch.org/fact-tank/2014/06/12/7-things-to-know-about-polarization-in-america/`.

40. Ellemers, N. The group self. *Science* **336**, 848–852 (2012).

41. Dehghani, M. *et al.* Sacred values and conflict over iran's nuclear program. *Judgment and Decision Making* **5**, 540 (2010).

42. Hvistendahl, M. Can predictive policing prevent crime before it happens. *Science Magazine* (2016). URL www.sciencemag.org/news/2016/09/can-predictive-policing-prevent-crime-it-happens.

43. Plato. *Plato's The Republic* (New York :Books, Inc., 1943).

44. Khomeini, I. & Khomeini, R. *Islamic government: governance of the jurist* (Alhoda UK, 2002).

45. Popper, K. R. *The poverty of historicism* (Psychology Press, 2002).

46. Ying, L., Hoover, J., Dehghani, M., Mooijman, M. & Ji, H. Acquiring background knowledge to improve moral value prediction. *ArXiv* (2017). URL arxiv.org/abs/1709.05467.

47. Hoover, J., Johnson, K. M., Dehghani, M. & Graham, J. Moral values coding guide. *PsyArXiv* (2017). URL psyarxiv.com/5dmgj.

48. Graham, J., Haidt, J. & Nosek, B. A. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology* **96**, 1029 (2009).

49. Haidt, J., Joseph, C. *et al.* The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. *The innate mind* **3**, 367–391 (2007).

50. Haidt, J., Graham, J. & Joseph, C. Above and below left–right: Ideological narratives and moral foundations. *Psychological Inquiry* **20**, 110–119 (2009).

51. Byrt, T., Bishop, J. & Carlin, J. B. Bias, prevalence and kappa. *Journal of clinical epidemiology* **46**, 423–429 (1993).

52. Sim, J. & Wright, C. C. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy* **85**, 257–268 (2005).

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

54. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2013). URL `http://www.R-project.org/`.

55. Gelman, A. & Hill, J. *Data analysis using regression and multilevelhierarchical models*, vol. 1 (Cambridge University Press New York, NY, USA, 2007).

56. Carpenter, B. *et al.* Stan: A probabilistic programming language. *Journal of Statistical Software* **20**, 1–37 (2016).

57. Buhrmester, M., Kwang, T. & Gosling, S. D. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science* **6**, 3–5 (2011).

58. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* **39**, 175–191 (2007).

59. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* **22**, 1359–1366 (2011).

60. Ohara, R. B. & Kotze, D. J. Do not log-transform count data. *Methods in Ecology and Evolution* **1**, 118–122 (2010).

Acknowledgements We would like to thank Antonio Damasio, Doug Medin, Scott Atran, Jonas Kaplan, Kingson Man Rumen Iliev, Sonya Sachdeva and UCSB's Psychology, Environment, and Public Policy group for their feedback on an earlier draft of this manuscript. This research was sponsored by the Army Research Lab. The content of this publication does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions Mooijman, Hoover and Dehghani developed the theory. Mooijman designed, ran and analyzed studies 2, 3 and 4. Hoover designed, ran and analyzed Study 1. Lin and Ji designed the text analysis method used to classify tweets. Mooijman, Hoover and Dehghani wrote the paper.

Competing Interests The authors declare that they have no competing interests.

Correspondence Correspondence regarding this article should be addressed to Marlon Mooijman (`marlon.mooijman@kellogg.northwestern.edu`), Joe Hoover (`jehoover@usc.edu`), or Morteza Dehghani (`mdehghan@usc.edu`).