Table of Contents

1.0 Twitter Field Studies

This section describes additional analyses for the Twitter field studies (Studies 1-3 in the main text). All data and code for Studies 1-3 is available at the following OSF link: https://osf.io/gtwsk/?view_only=10eda3c3f5924c399439e2102e18ae97. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study in our preregistrations of Study 2 (https://osf.io/ud5bc) and Study 3 (https://osf.io/qmnyb).

1.1 Examining the validity of DOC's outrage classifications of feelings

In our field study, we collected messages that were classified by DOC as expressing outrage (probability greater than 95% according to DOC's output), and also those that were classified as not expressing outrage (probability lower than 50% according to DOC's output). Thus, we could use the authors' reports of outrage to test the validity of the classifications of DOC in terms of emotional experience. Analyzing all tweets collected in the Author Phase across studies, a Wilcoxon signed-rank test revealed that authors' self-reports of outrage were significantly greater for tweets classified by DOC as outrage ($M = 4.31$) compared to those that were classified as non-outrage ($M = 3.66$), $W = 11420$, $p = .009$, $d = .28$. These results suggest that DOC's classifications tracked outrage *feelings* reported by authors, even though DOC was trained based on identifying linguistic features of outrage *expression* in text.

1.2 Analysis of discrepancy between author and observer ratings

As a robustness test to the analyses reported in the main text, we tested the overperception hypothesis by comparing the authors' self-reported moral outrage for each tweet to the mean observer judgments of authors' moral outrage in the same tweets (setting up a group means analysis instead of using the multi-level model). As the distributions of outrage ratings were not normal, we conducted a Wilcoxon signed-rank test to compare the mean differences between author and observer ratings. Consistent with findings reported in the main text, we found that observes judged authors to be significantly more outraged than the authors themselves reported (Study 1, $M_{diff} = 0.58$, $p = .006$, $d = .25$; Study 2, $M_{diff} = 0.58$, $p < .001$, $d = .28$; Study 3 $M_{diff} = 0.63$, $p < .001$, $d = .30$). Also consistent with analyses reported in the main text, the observed discrepancy between authors and observers was selective to moral outrage. There was no evidence for a discrepancy between authors' and observers' happiness ratings (Study 1, $M_{diff} = -0.13$, $p = .002$, $d = .07$; Study 2, $M_{diff} = -0.17$, $p = .653$, $d = .09$), see **Fig. 2**. Furthermore, the discrepancy between author reports and observer judgments was significantly larger for outrage compared to happiness (Study 1, $M_{diff} = 0.72$, $p = .015$, $d = .21$; Study 2, $M_{diff} = .75$, $p < .001$, $d = .24$).

As another robustness test, we ran our main multi-level models reported in the main text while also adjusting for each observers' happiness judgments. The overperception finding was still significant in these models: Study 1, $b = 0.52$, $p = .029$; Study 2, $b = 0.51$, $p = .003$; Study 3, $b = 0.60$, $p < .001$, suggesting that variance in happiness ratings does not explain the overperception finding.

Further exploring accuracy and bias

The accuracy of group meta-perceptions can also be understood more finely as a combination of group-level over/underestimation of values (as we tested above), but also the rank-order accuracy across targets and judgments [1]. We attempted to follow a model suggested in previous work [1] to tease these two components of accuracy apart, but the model would not converge, perhaps due to having a limited number of targets or judgments. However, these models did suggest that although observers overestimated author outrage at the group-level, they were still tracking some signal of outrage accurately. For instance, in a separate analysis, we found that observers' judgments were significantly and positively correlated with authors' reports of outrage in Study 1: $r(131) = .23$, $p = .007$; Study 2: $r(193) = .48$, $p < .001$; Study 3: $r(195) = .47$, $p < .001$. These results suggest that observers were generally tracking variation in outrage expression, but they overperceived the intensity.

Examining lack of confidence as a source of overperception

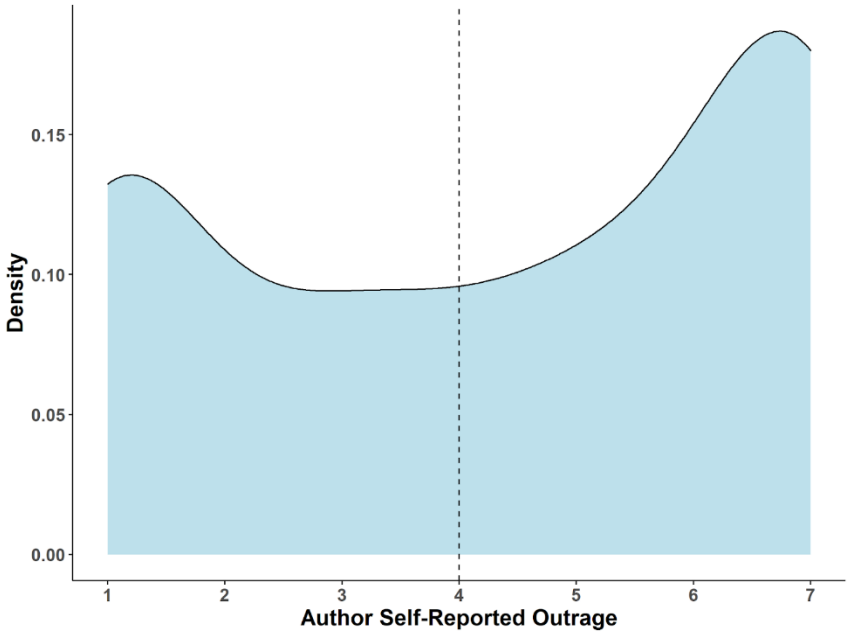
Another question is whether observers' overperceptions were driven by the fact that they were unconfident with their judgments, and so they simply guessed around the midpoint for many tweets where the author reported low amounts of outrage (thus driving the observer group mean high artificially). However, observer judgments did not appear to be driven by low confidence: the average confidence ratings across all tweets was above the midpoint on the 7-point Likert-type confidence rating scale (Study 1: $M = 5.50$, range = 3.76 – 6.63; Study 2: $M = 5.49$, range = 4.18 – 6.18; Study 3: $M = 5.60$, range = 3.43 – 7.00).

1.3 Is overperception of outrage driven by authors or observers?

One question is whether there is evidence that authors drive overperception of outrage by overexpressing more outrage than they feel, or whether observers drive overperception of outrage by perceiving outrage in a biased manner. There are theoretical reasons to believe that both are driving the discrepancy between author reports and observer judgments that we document in our field studies (see main text, discussion), yet our studies were not necessarily set up to precisely disentangle author vs. observer effects. Here, we present analyses that provide indirect evidence for author vs. observer effects.

Rather than overperception by observers, it could be the case that our results are purely driven by authors under-reporting their high outrage feelings even though they are expressing high outrage. Below, we examine three potential pattern of results that would be expected if our results were purely driven by authors under-reporting their high outrage feelings.

First, if our results were purely driven by authors under-reporting their outrage, we should find that authors are not very likely to report high outrage in our field studies. To the contrary, across our field studies we find that for those messages DOC identified as expressing moral outrage, 60% of message authors self-reported being somewhat morally outraged or greater, see **Supplementary Figure 1** below:

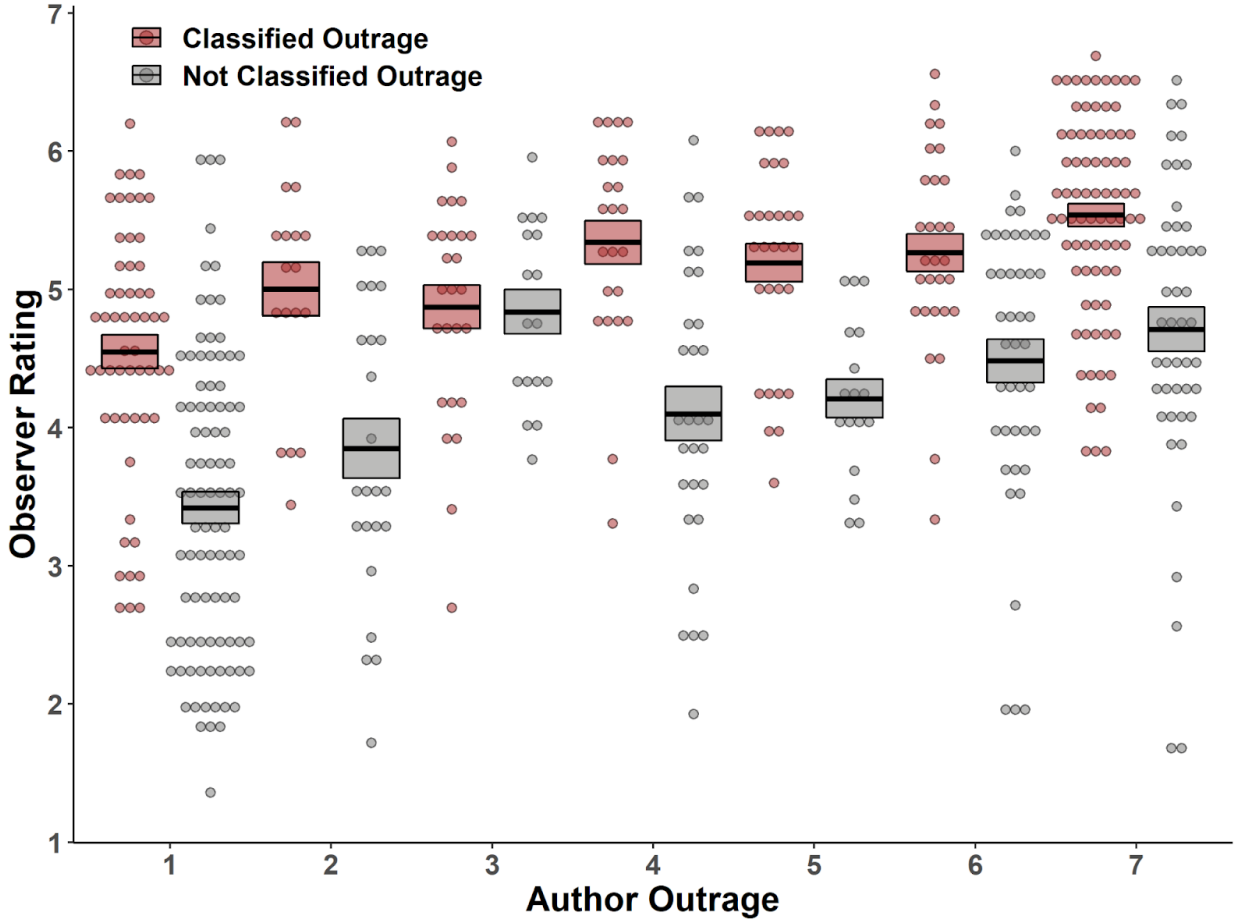


Supplementary Figure 1. Distribution of authors' self-reported moral outrage for messages classified by DOC as containing moral outrage expression. The dotted line represents a rating of "4", which indicates that the author self-reported that they were "somewhat" morally outraged. The scale ranges from 1 (not at all outraged) to 7 (very outraged).

Second, if our results are purely driven by authors under-reporting high outrage feelings even though they express high outrage, we should see evidence that discrepancies between author and observer ratings only occur for the messages identified by DOC as expressing outrage. DOC is validated to detect outrage expression signals, so those messages DOC *does not* classify as outrage are much less likely to have true signals of outrage observers are picking up on. However, we see evidence that observers are systematically reporting greater outrage for *both* messages that DOC detected outrage expression and messages that DOC did not detect outrage expression, see **Supplementary Figure 2**.

Third, our results could be driven by a combination of a scale response artifact and authors who mostly report low outrage: on average observers simply tend to judge messages at the midpoint of the scale (a "4") across the board, if most message authors report low outrage, and then on average it would appear that observers are overperceiving the outrage overall. This

explanation is ruled out for two reasons. First, **Supplementary Figure 2** clearly shows that observers are not simply reporting "4" on average. In fact, observers are sensitive to increases in authors' self-reported outrage on average even with no access to the self-report. This finding is also consistent with the significant correlation between author self-report and observer ratings, despite overperception occurring on the average, see SI Appendix, Section 1.2. Second, **Supplementary Figure 1** already established that authors are not simply reporting low level of outrage that would be required to drive this explanation (in fact, the opposite is true). We also note that when authors report a "7" for their outrage feelings, observers' overperception cannot be measured because the highest they can judge is a "7" on the same scale.



Supplementary Figure 2. Observer outrage ratings plotted as a function of author outrage ratings. Red dots and bars represent ratings for messages that DOC classified as containing outrage expression. Gray dots and bars represent ratings for messages that DOC classified as not containing outrage expression. Box plots represent mean +/- 1 standard error of the mean. Overperception is indicated when observer ratings are higher than author ratings of outrage. *N* tweets = 333.

Next, we examined author and observer individual differences to test whether they might be correlated with overperception of outrage. We first tested whether the ideological extremity of authors would predict the mean levels of overperception for their messages, reasoning that more ideologically extreme authors (who tend to be more strongly identified with their political party) might have stronger motivations to express high outrage for reasons related to group identity concerns (see [2] for detailed discussion). For message authors (all of whom participated in our field study), we estimated their political ideology using the 'tweetscores' package in *R* 3.6.3 [3]. Ideological extremity was determined by computing the absolute value of the continuous values produced by the tweetscores estimates. Examining all message authors who we were able to estimate ideological extremity values ($N = 299$), we observed no significant relationship between ideological extremity of the author and the extent to which their messages were overperceived, $r(297) = -.09$, $p = .673$.

We also tested whether authors social reinforcement history (number of likes / shares received on average for tweeting outrage across the users' history) and audience size (follower count) was associated with overperception. These variables may be proxies for either habitual outrage expression or motivations to express outrage performatively. However, we found no association between reinforcement history and outrage expression, $r(249) = -.02$, $p = .763$, nor audience size and overperception of outrage $r(249) = -.00$, $p = .970$.

Regarding observers, we tested whether individual differences in ideological extremity and political identification strength were associated with the extent to which observers tended to

overperceive outrage in the messages they saw. Examining data across Studies 1-3, we observed no significant relationship between ideological extremity of observers and their magnitude of overperception, $r(582) = .02$, $p = .555$. There was also no relationship between partisan identity strength and their magnitude of overperception of outrage, $r(582) = .05$, $p = .260$.

As reported in the main text, one observer characteristic that was associated with overperception in Studies 1-3 was observers' daily political social media use. This relationship was still significant when statistically adjusting for observers' ideological extremity and partisan identity strength in a linear regression model, Studies 1-2: $b = .17$, $p = .009$; Study 3: $b = .16$, $p = .002$. It was also still significant when statistically adjusting for observers' tendency to overperceive happiness, Studies 1-2: $b = .21$, $p = .002$; Study 3: $b = .19$, $p < .001$.

1.4 Analyzing only frequent social media users

One concern is that the authors who were recruited on Twitter are more active on social media than observers recruited on Prolific, and thus they have a different understanding of outrage expression on social media. This concern is mitigated by the fact that we only recruited observers on Prolific who were active social media users. To examine this more conservatively, we reproduced the overperception analysis using only observers who are very active on social media (use social media at least 4-6 times per week). We found that the overperception finding held for this analysis in all studies (Study 1: $b = 0.59$, $p = .001$, Study 2: $b = 0.60$, $p = .013$, Study 3: $b = 0.62$, $p < .001$), suggesting that discrepancies between author and observer social media activity does not explain the overperception finding.

1.5 How generalizable are results given attrition from our DM method?

One question is how generalizable our findings are given that there was attrition with our DM method. To examine this issue carefully, we ask the following questions:

1. What are some relevant dimensions on which our message authors might differ from other Twitter users in ways that could threaten the generalizability of our results?

2. What are the relevant comparison groups of Twitter users against which we should compare our message authors to examine generalizability?

For the first question, our reviewers suggested several relevant comparison metrics: (a) overall amount of outrage expression, (b) political extremity, (c) overall Twitter engagement/activity, and (d) number of followers.

Metric (a) is relevant because if the field study message authors are systematically more likely to express outrage than comparison groups, then they may be so accustomed to expressing outrage that their *expression* is always high even though they may not always *feel* high outrage. This would mean that our field study authors are more likely to produce overperception effects than the comparison groups, because observers would perceive high outrage in their tweets but these authors would not always report corresponding feelings of outrage. On the other hand, if the field study message authors are systematically less likely to express outrage than comparison groups, then this would suggest that our field studies represent a conservative test of our overperception hypothesis. In other words, lower rates of outrage expression reduces the range of potential discrepancies between outrage expression and outrage feelings and thus reduces the

chances for observing discrepancies between author self-reports of outrage and observers' perceptions.

Metric (b) is relevant because if field study message authors are systematically less politically extreme than comparison groups, they may be less emotionally charged or invested in the political topics they tweet about. Thus, even when they *express* outrage about these topics, they may *feel* less outrage relative to other users who are more politically extreme.

Metrics (c) and (d) are relevant because if field study message authors have disproportionately high Twitter engagement or audience size, they may be more likely to be using Twitter for performative/PR/professional reasons and thus more inclined to overstate outrage publicly in the pursuit of clicks. We consider all four of these metrics in the analyses reported below.

The second question concerns identifying the population(s) for which we are seeking to generalize our results. On theoretical grounds, we do not think it is appropriate to compare our message authors to all Twitter users, because the phenomenon we study concerns the overperception of outrage in messages that display on a Twitter newsfeed, and it is well established that a minority of Twitter users generate the majority of political Twitter content (see citations 21-25 in the main text of the manuscript). Thus, it is more appropriate to determine how our message authors compare to narrower subgroups of Twitter participants who actually produce content about the political topics we study.

Below, we report analyses comparing our message authors on metrics (a)-(d) above with 2 relevant comparison groups:

1. **Twitter users whom we contacted during our study but did not respond to our DM**. This comparison group of "no reply users" directly addresses the attrition concern

because these users were tweeting about the same topics as our message authors, and recruited to our study in the same way, but chose not to respond to our DM.

2. **Twitter users who tweeted about a different politically charged topic**. This comparison group addresses both the attrition concern as well as broader questions about generalizability beyond the specific political topics included in the current study. We ask whether our message authors differ on key metrics from a set of Twitter users who were not selected into a study via DM and tweeted about a different politically charged topic (the Brett Kavanaugh Supreme Court confirmation hearings in 2018).

The results of these new analyses are summarized in the table below:

Group	Mean Outrage %	Ideological Extremity	Median Tweets per Year	Median Follower Count
Field Study Authors	23% (10%)	1.29 (0.61)	803 (624)	376
Users who did not respond to Field Study DM	30% (14%)	1.21 (0. 71)	805 (853)	1004
Users who tweeted about different politically charged topic	26% (12%)	1.50 (0.53)	512 (568)	182

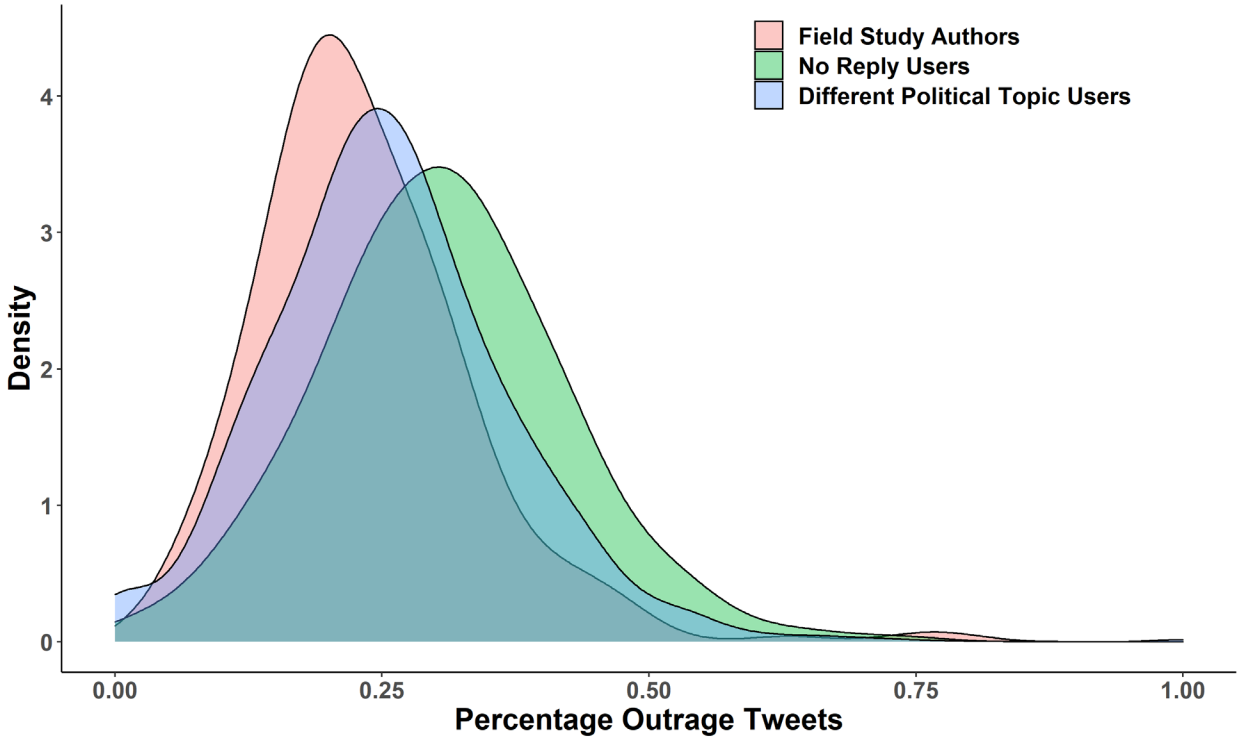
Supplementary Table 1. Characteristics of Twitter users from the field studies in this manuscript, a random sample of 600 users who did not reply to our DM and a sample of 3,669 users who tweeted about a different contentious political topic (Brett Kavanaugh confirmation hearings; Brady et al., 2021, *Science Advances*). Mean outrage % displays the mean percentage of users' tweet history estimated to contain outrage expression as classified by DOC, and standard deviations of the percentage are displayed in parenthesis. Ideological extremity displays the mean and standard deviation of a continuous estimate of ideological extremity (Barbera, 2015), where greater values indicate greater ideological extremity. Median tweets per year displays the median of the mean number of tweets users tweeted per year, parenthesis display the standard deviation of mean tweets per year. Median follower count displays the median followers count for users, estimated at the time of data collection. Because outliers on follower

count highly skew the standard deviation, they are not displayed for follower count, but density plots show the distribution in Fig. S4.

Field study authors do not express more outrage than comparison groups

To test how our field study message authors compare to the comparison groups in terms of outrage expression, we collected the entire Tweet history of message authors in our field studies (N_{tweets} = 710,240) and for each message author we computed (1) the mean percentage of their past tweets that contained outrage expressions according to our Digital Outrage Classifier (DOC).

To form our comparison group of users who did not respond to our DM, we took a random sample of 700 users from our database of users who did not respond. This number represents the highest number of users for which it was computationally feasible to collect their entire tweet histories and compute their outrage expression in a timely manner. Our final dataset consisted of N = 2.03 million tweets from the 700 users. To form our comparison group of users who tweeted about a different contentious political topic, we used an existing dataset of users from a previous study (Brady et al., 2021, *Science Advances*, Study 1) consisting of 3,669 users and 6.1 million tweets. We used DOC to compute the mean percentage of tweets that contained outrage expressions for all users in the two comparison sets, and the same political ideology estimating technique as for our field study authors. **Supplementary Table 1** and **Supplementary Figure 3** compare the mean outrage expression of each group.



Supplementary Figure 3. Density plot comparing the mean percentage of outrage in users' Tweets among field study authors (n = 710,240), users who were contacted but did not reply to our field study DMs (n = 2,037,606) and users who tweeted about a different political topic than those appearing in our field studies (n = 6,104,194).

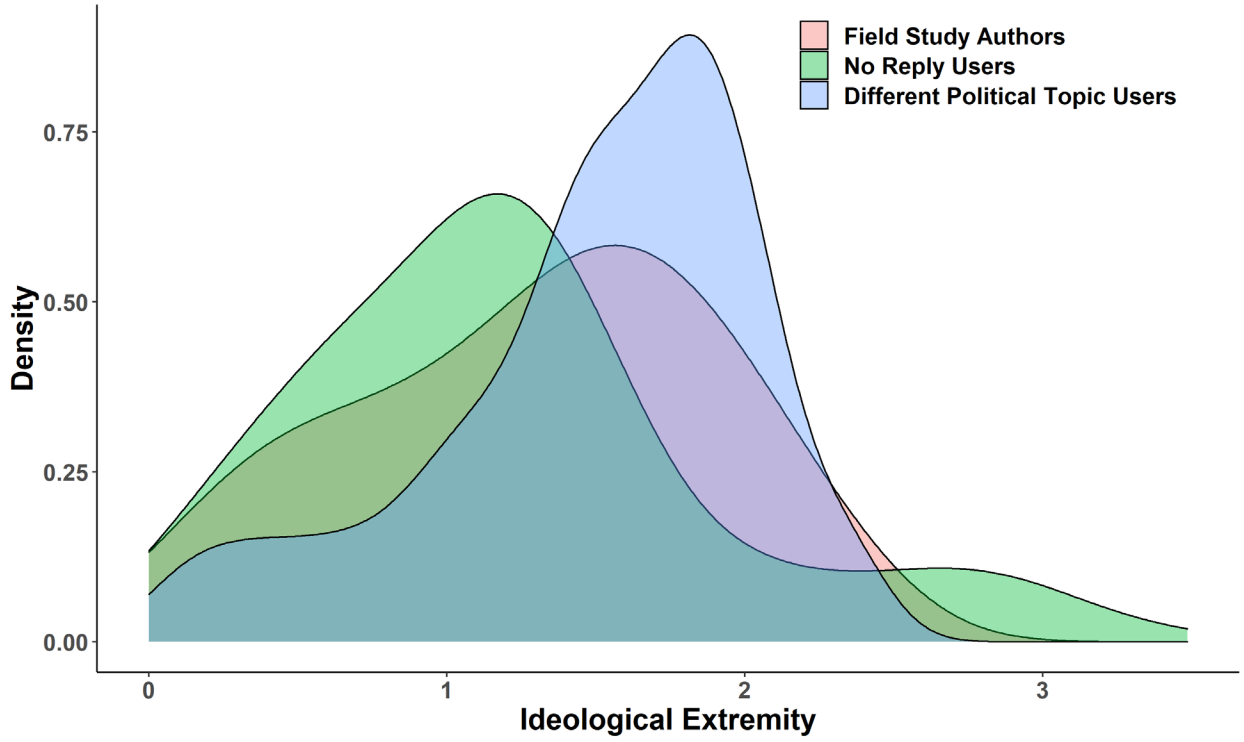
As displayed in **Supplementary Table 1** and **Supplementary Figure 3**, we find that the amount of outrage expressed in the field study authors' timelines are not significantly higher than either comparison group, and generally all groups are within the range of each other's outrage expression when comparing means and standard deviations.

Because outrage expression of our field study authors was not significantly higher than the 'no reply users', we have no evidence that attrition effects can explain our overperception findings. In other words, the data suggest our field study authors are not more likely to overexpress or under-report outrage compared to users who did not respond to our DM. Because outrage was not higher than (and were very similar to) users who tweeted about a different political topic, we also have evidence that our results should generalize well to other politically

active users (our population of interest). Because field study authors expressed slightly less outrage overall than comparison groups, we believe that this participant group presents a conservative test of our hypothesis. Nonetheless, because field study authors expressed slightly less outrage, data collection with a larger number of users is required to ensure that our results generalize to users who express more outrage overall. Importantly, however, we find no evidence that variation in frequency of outrage expression in study authors' timelines predict the extent to which their outrage is overperceived, $r(249) = -.06$, $p = .363$.

Field study authors do not systematically differ from comparison groups on political extremity

To compare political extremity, we estimated the political ideology of each user in all groups on a continuous dimension based on the political accounts that they follower on Twitter (using the 'tweetscores' package; Barbera, 2015), see **Supplementary Table 1** and **Supplementary Figure 4**. We found that field study authors were slightly more extreme than 'no reply users' but slightly less extreme than users who tweeted about a different political topic, although every group falls within one standard deviation from each other. Overall, the field study authors fall in between comparison groups regarding their political extremity, suggesting that they are not systematically less extreme than comparison groups. Furthermore, we find no evidence that the ideological extremity of field study authors predict the extent to which their outrage is overperceived, $r(297) = -.09$, $p = .673$, see SI Appendix Section 1.3



Supplementary Figure 4. Density plot comparing the mean political ideological extremity of field study authors (n = 710,240), users who were contacted but did not reply to our field study DMs (n = 2,037,606) and users who tweeted about a different political topic than those appearing in our field studies (n = 6,104,194). Political ideology was estimated using the 'tweetscores' package (Barbera et al., 2015).

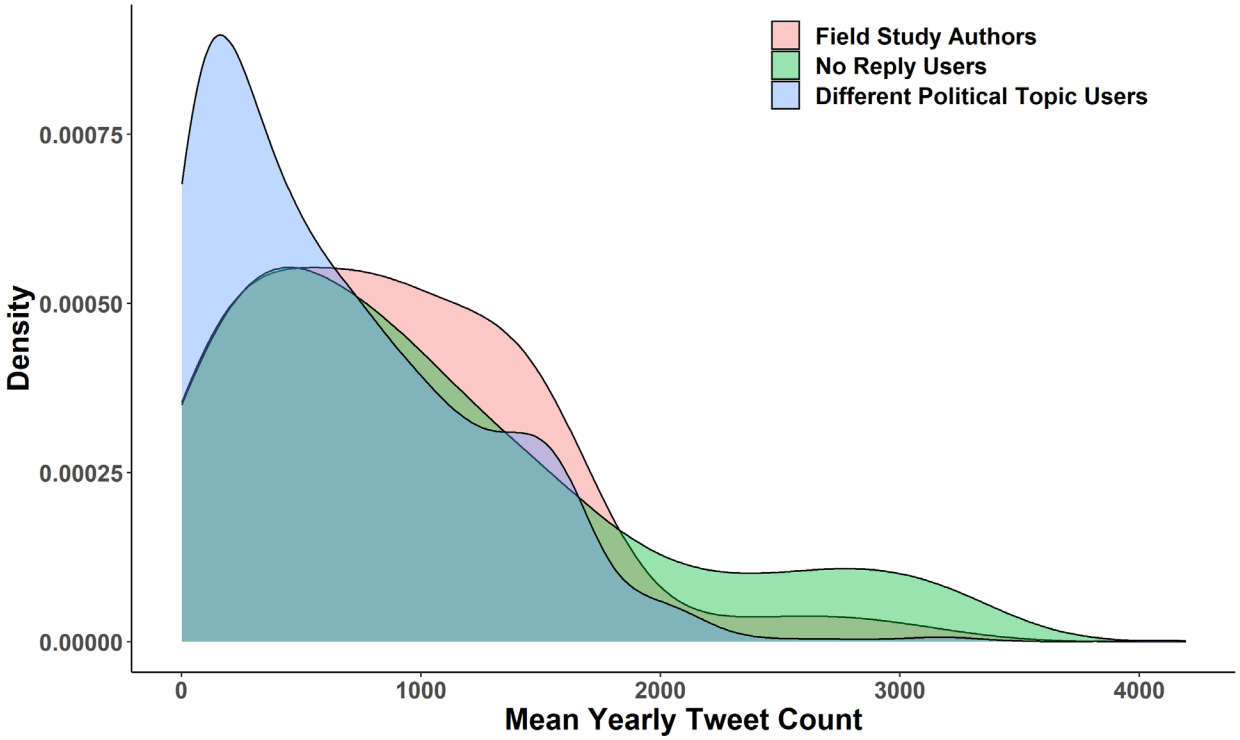
Field study authors do not systematically differ from comparison groups in tweet counts per year and followers

We measured tweet counts per year because variation in tweet counts can represent the amount of engagement users have with Twitter. We find that field study authors do not systematically differ from comparison groups in tweets per year: they do not tweet significantly differently than 'no reply' users, but tweet significantly more than users who tweet about a different political topic (although the yearly tweet activity is well within one standard deviation range of users who tweet about a different political topic).

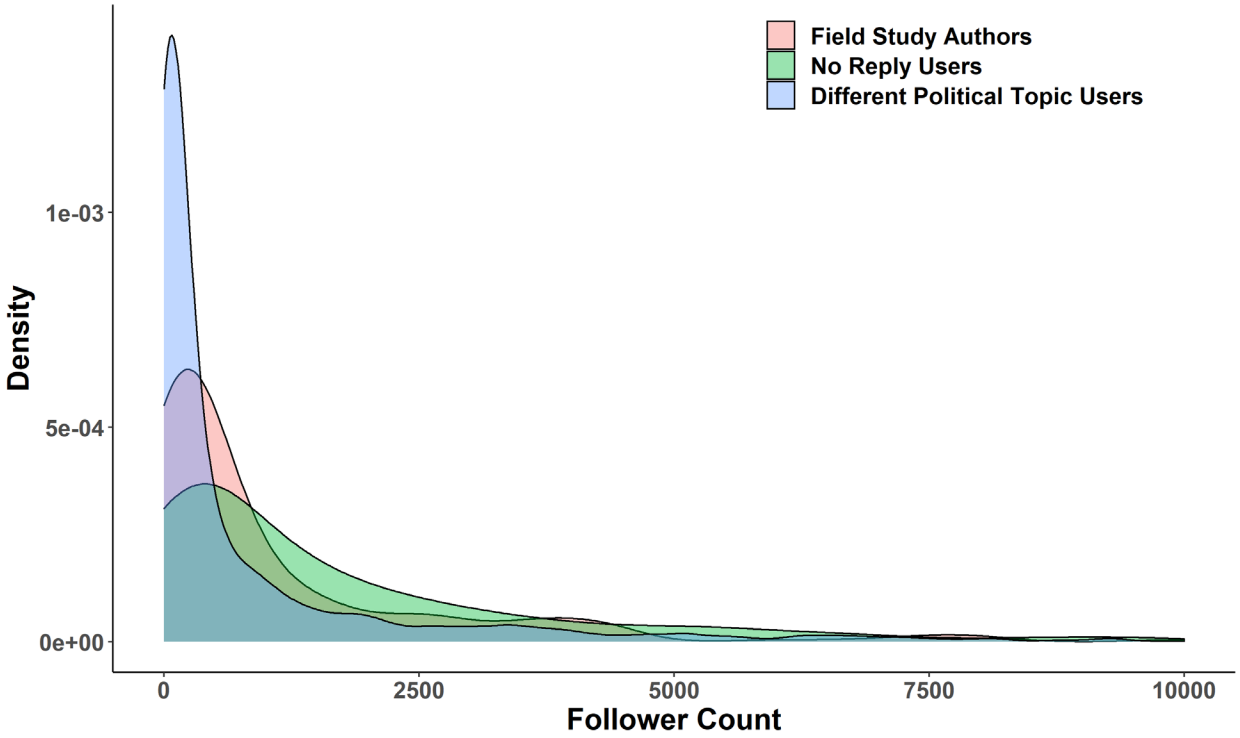
Because field study authors tweet the same amount as 'no reply users', we can rule out the idea that attrition led to the selection of authors who tweet more for engagement and PR

reasons. Because field study authors are within the standard deviation range of tweet activity compared to users who tweet about a different political topic, it is highly unlikely field study authors engage with Twitter to an extent that impacts the generalization of our overperception findings to other politically active users. In fact, we find that no relationship between yearly tweet count and the extent to which authors are overperceived within our field studies, $r(249) = .07$, $p = .283$. See **Supplementary Figure 5** for density plots of yearly tweet counts comparing each group of users.

We measured follower counts because variation in follower counts is a proxy for audience size. If field study authors have greater followers / audience size, it could raise the possibility that they are more strongly motivated to express outrage for performative / PR reasons. We find that field study authors have significantly fewer followers than 'no reply' users, ruling out the idea that attrition effects lead to the selection of authors who are especially motivated to express outrage for performative reasons due to larger audience size. We find that field study authors have significantly more followers than users who tweeted about a different political topic, which could suggest that field study authors may be impacted by their larger audience size compared to other political users. However, we found that follower count did not predict the overperception finding among field study authors, $r(249) = -.04$, $p = .970$. See **Supplementary Figure 6** for density plots of follower counts comparing each group of users.



Supplementary Figure 5. Density plot comparing the mean yearly tweet count of field study authors (*n* = 710,240), users who were contacted but did not reply to our field study DMs (*n* = 2,037,606), and users who tweeted about a different political topic than those appearing in our field studies (*n* = 6,104,194).

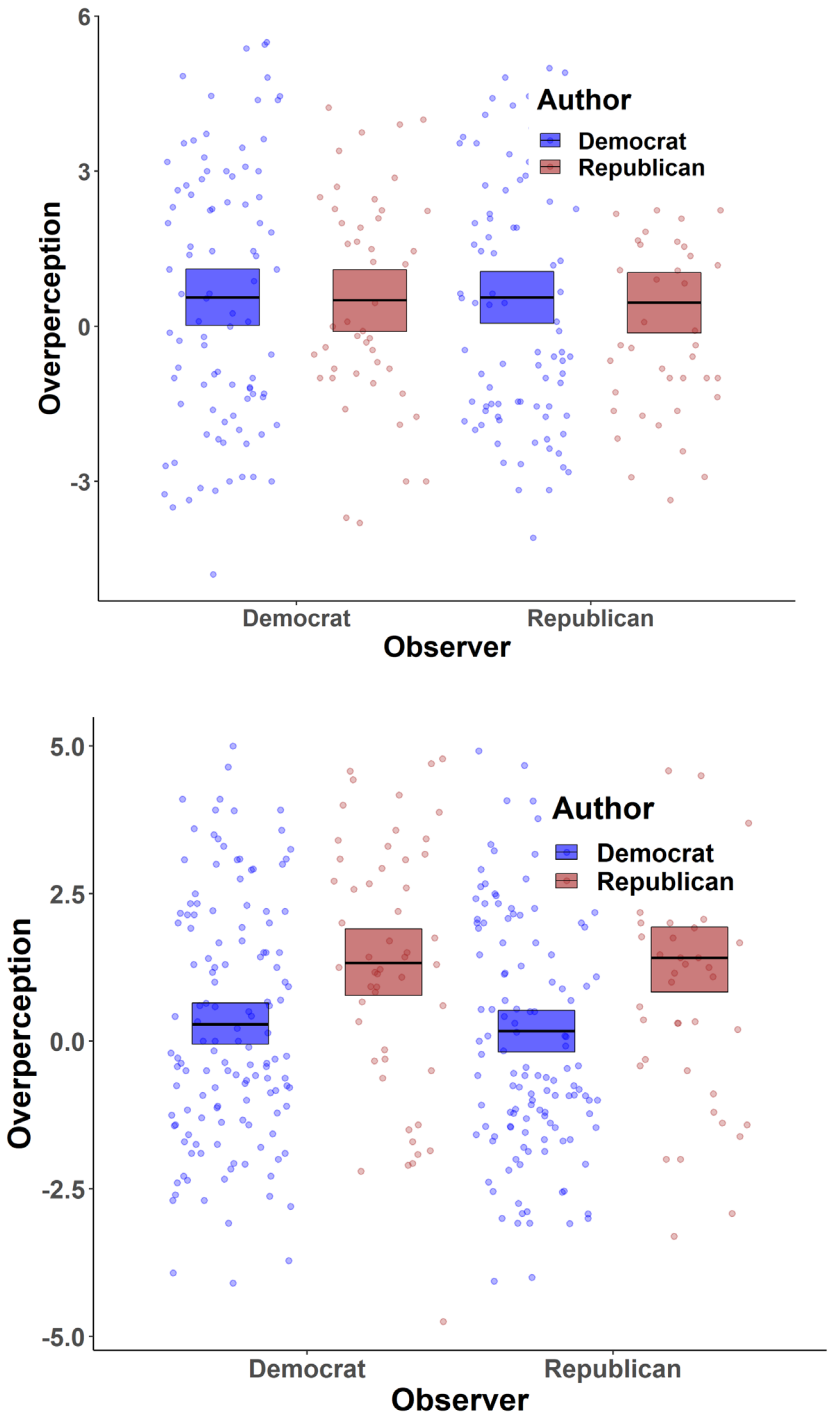


Supplementary Figure 6. Density plot comparing follower counts of field study authors (n = 710,240), users who were contacted but did not reply to our field study DMs (n = 2,037,606), and users who tweeted about a different political topic than those appearing in our field studies (n = 6,104,194). Outlier users with greater than 10,000 followers are trimmed from the plot for readability.

1.6 Ingroup vs outgroup differences in overperception

In Studies 1-3, observers made judgments of both ingroup and outgroup Twitter messages. Thus, we examined whether observers tended to overperceive ingroup vs. outgroup messages differently. Author partisanship was determined based on consensus judgment of the position being argued in the Tweet text, observer partisanship was determined by their self-reported political party affiliation. Here we analyze Study 1 and 2 since they were the only studies for which we obtained author political party estimates. Results were inconsistent, as in Study 1 we did not find significant differences in the mean overperception of outrage for ingroup vs outgroup members, nor in Study 2. However, in Study 2 we found that Republican authors

were overperceived more than Democrat authors (by both Republican and Democrat observers). **Supplementary Figure 7** displays results.



Supplementary Figure 7. Comparison of overperception based on ingroup vs. outgroup for Study 1 (top) and Study 2 (bottom). The Y axis represents mean overperception (observer judgment – author self-report) of moral outrage for each political party. The left side of the graph displays Democrat observers, the right side of the graph displays Republican observers. Blue bars represent Democrat authors being judged, red bars represent Republican authors being judged. Box plots represent mean +/- 1 standard error of the mean. *N* messages = 333.

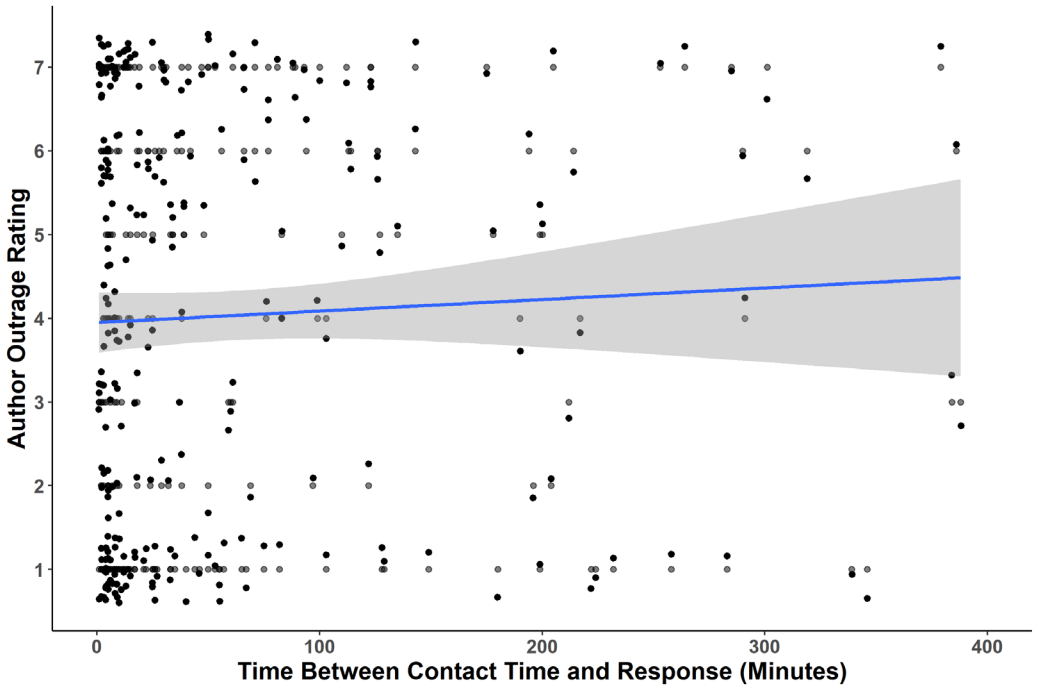
1.7 Differences in overperception result as a function of author partisanship

One question is whether the authors' political partisanship impacted the extent to which a message was overperceived by observers. To examine this, we entered the source of judgments (author vs observer), author partisanship and their interaction in our main multilevel model. In Study 1, we found no evidence that author partisanship interacted with the overperception estimate, $b = 0.04$, $p = .793$. In Studies 2 and 3 (which used the same author targets), a significant interaction found that Republican message authors were overperceived more than Democrat message authors, Study 2: $b = 0.53$, $p < .001$; Study 3: $b = 0.64$, $p < .001$. Greater overperception of Republican authors was found both for Republican observers and Democrat observers, see SI Appendix, Section 1.5 above.

1.8 Do author self-reports of moral outrage vary based on response time?

One possibility is that our overperception results are created by a confound: authors who responded later to our DM may have reported lower outrage than their message entails, because they were less likely to be feeling the same level of moral outrage as more time has passed. To test this possibility, for all messages authors contacted in each study, we calculated the difference in minutes between the time our DM was sent to the author and when they responded. We then correlated this time difference with their self-reported moral outrage value. We found no evidence that authors who responded later were likely to self-report lower moral outrage values, $r(335) = -.03$, $p = .559$. We also correlated response time with the tendency for an

authors' message to be overperceived by observers, but again found no correlation, $r(323) = .08$, $p = .162$. **Supplementary Figure 8** displays authors' outrage ratings as a function of author response time.



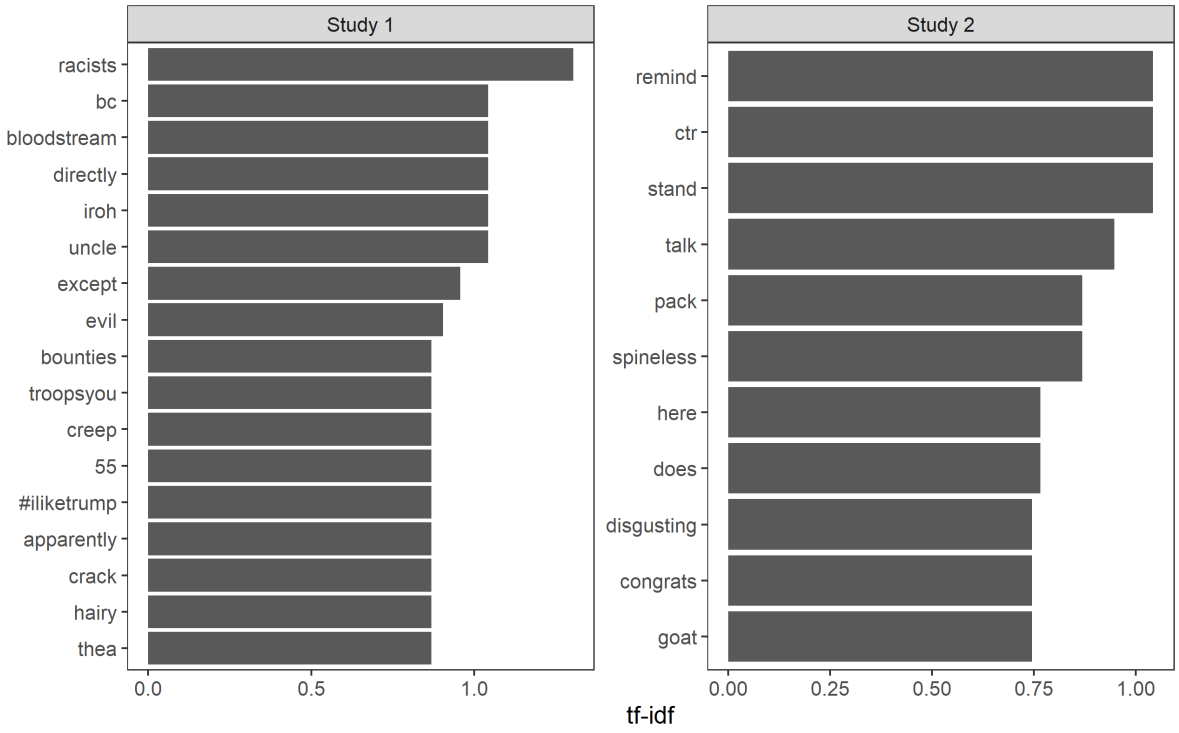
Supplementary Figure 8. Author moral outrage rating as a function of time between contact time and response in minutes. Error bands represent 95% CI's on linear model predictions.

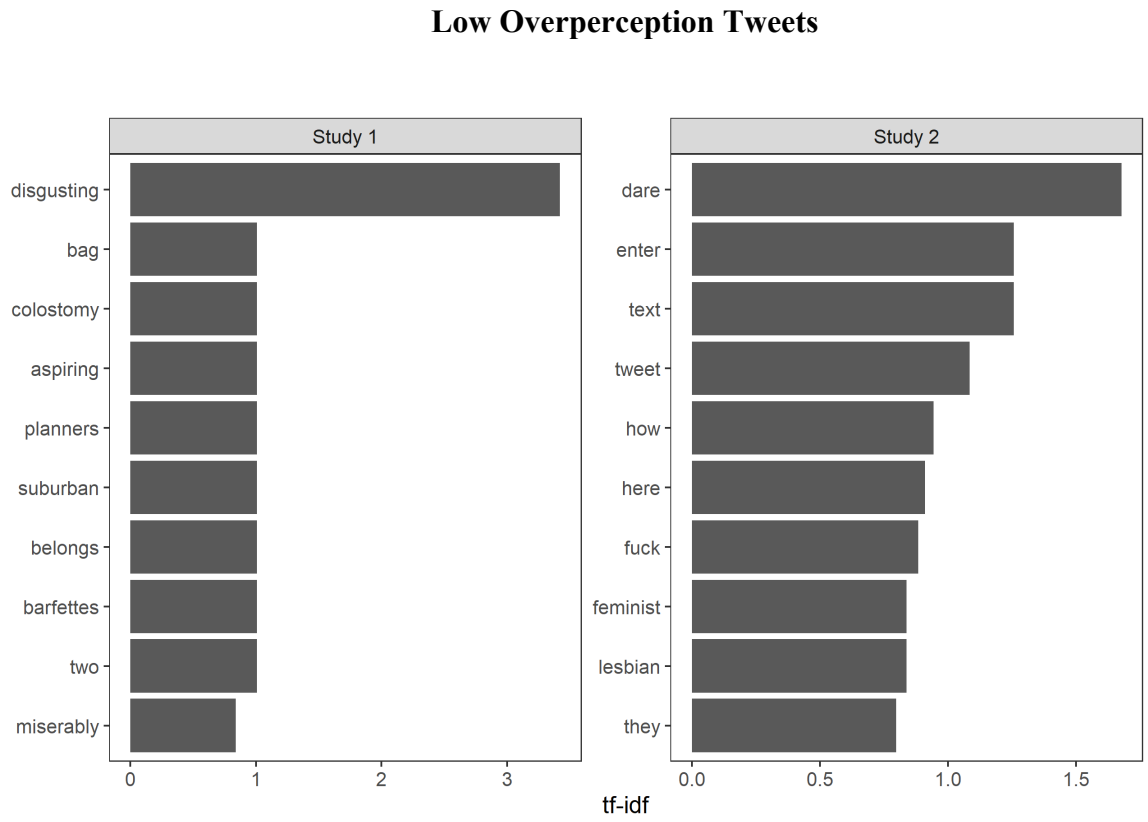
1.9 Linguistic features associated with overperception in our studies

In an exploratory analysis, we conducted a tf-idf analysis in order to determine which words were most "important" for messages that tend to the most overperceived regarding their moral outrage expression (i.e., the most frequently used words accounting for the base rate of word usage), see **Supplementary Figure 9**. However, we note that these words should not be interpreted as those that universally trigger overperception of moral outrage, because they are highly contextually bound to the topics that were included in our study. It is interesting to note

that several words associated with moral and emotional expression appear as key text features that predict whether a tweet is overperceived in these studies including words such as, “racist”, “evil”, “spineless” and “disgusting” (although “disgusting” also serves as a clear signal of outrage since it also appears a key textual feature in low overperception tweets).

High Overperception Tweets





Supplementary Figure 9. Tf-idf analysis of words that were most "important" for high overperception tweets (i.e., the most frequently used words accounting for the base rate of word usage), and words that were most "important" for low overperception tweets.

1.10 Statistically controlling for age in social media use / overperception models

In an exploratory analysis, we ran the models from the main text examining the association between political social media use and overperception of outrage, this time also controlling for observer age in the model. For Studies 1-2 and Study 3, we found that the association between political social media use and overperception of outrage remained significant when controlling for age, Studies 1-2, $b = 0.17$, $p = .010$, Study 3: $b = 0.16$, $p = .002$. Age was not a significant predictor of overperception of outrage in these models: Studies 1-2, $b = -0.01$, $p = .823$, Study 3: $b = 0.01$, $p = .857$.

1.11 Log-transforming political social media use

As the political social media use variable was positively skewed, we log-transformed the variable (political social media use + 1 to account for 0 values) and re-ran our correlation analysis reported in the main text. These analyses replicated the correlation between political social use and overperception of moral outrage, Studies 1/2: $r(222) = .20$, $p = .003$; Study 3: $r(248) = .20$, $p = .002$.

1.12 Full covariate models for political social media use analysis

We ran a model testing the relationship between political social media use and overperception of moral outrage while statistically controlling for partisan identity strength, political ideology and overperception of happiness entered as competing covariates. These models replicated the significant association between political social media use and overperception of moral outrage, Studies 1/2: $b = .21$, $p = .003$; Study 3: $b = .17$, $p = .002$.

1.13 Further exploring overperception of happiness

Although we did not find evidence for overperception of happiness on average, one question is whether overperception of happiness occurs for those tweets where authors reported high happiness. If overperception does not occur, it would further bolster the claim that overperception is specific to moral outrage and not happiness. Although we did not specifically select for high vs. low happiness tweets in our selection of authors (as we did for outrage), we exploited natural variation of happiness self-reports in our dataset to run this exploratory analysis. Across all studies, we find no evidence for overperception of happiness for high happiness tweets, and in fact we find evidence of underperception, Study 1: $b = -2.34$, $p < .001$;

Study 2: b = -2.64, p < .001; Study 3: b = -2.68, p < .001. These results further support the idea of a negativity biases in emotion perception online and demonstrate that overperception only occurred for outrage in our studies.

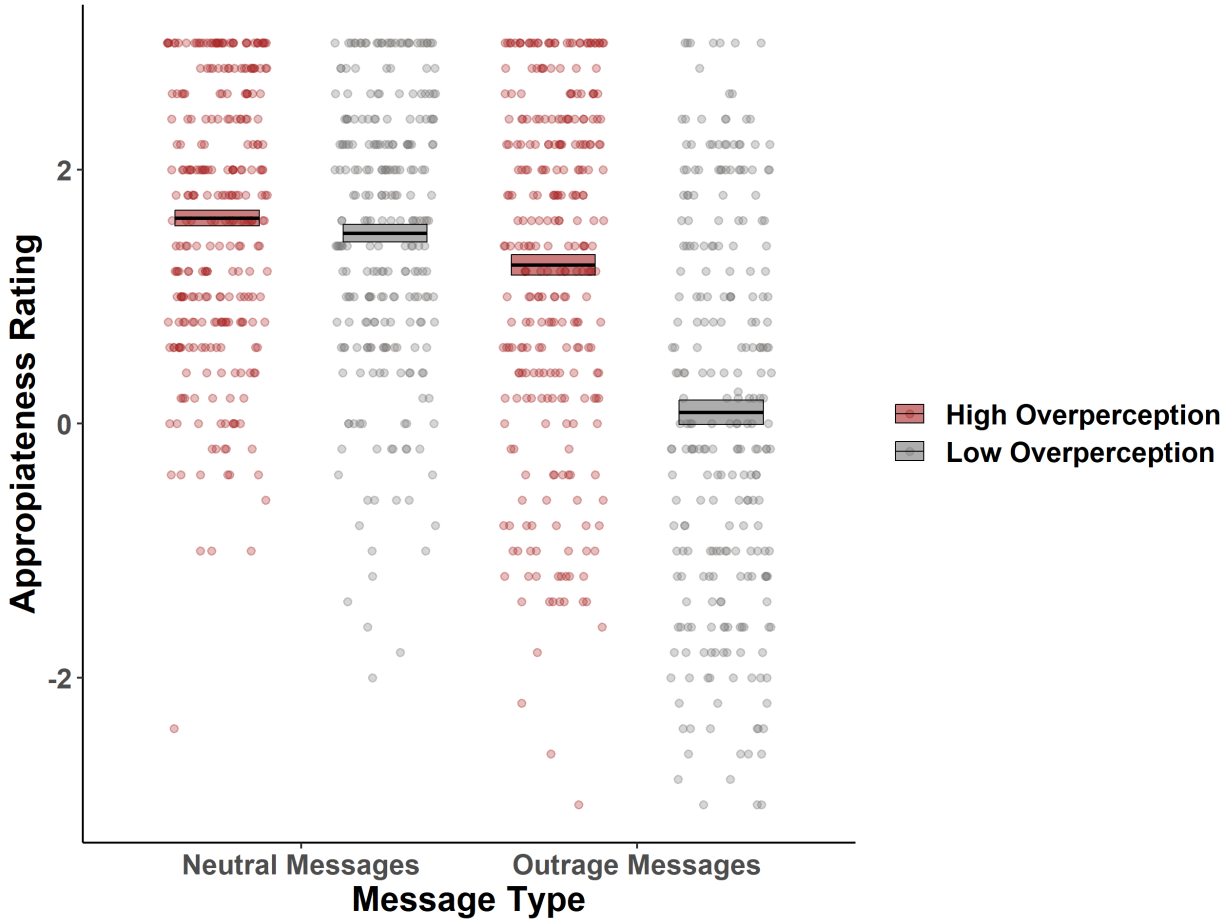
2.0 Behavioral Experiments

This section describes additional analyses for the behavioral experiments (Studies 4 and 5 in the main text). All data and code for Studies 4-5 is available at the following OSF link: https://osf.io/gtwsk/. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study in our preregistrations of Study 4 (https://osf.io/mjftk) and Study 5 (https://osf.io/sxtah).

2.1 Examining raw norm ratings (Study 5)

The main text reports the difference score between appropriateness ratings for outrage vs. neutral messages viewed by participants. We used the difference score (pre-registered analysis plan) to reveal how appropriate participants in each condition thought outrage messages were while adjusting for their baseline appropriateness ratings derived from their ratings of neutral messages. Below, we show raw ratings that plot both outrage and neutral message ratings. **Supplementary Figure 10-11** below show ratings for both Republican and Democrat participants (participants only viewed in-party messages, thus the stimuli were different).

Supplementary Figure 10. Appropriateness ratings in response to a norm task as a function of condition and type of message (Republican participants). Participants were asked to judge how appropriate it would be to post neutral vs. outrage messages to the social network they viewed depending on their condition. In the high overperception condition participants viewed messages in their newsfeed that tended to judge as containing more outrage than the author reported, as determined by Studies 1 and 2. In the low overperception condition participants viewed messages in their newsfeed that tended to be judged as containing similar outrage compared to the author's self-report. Box plots represent mean +/- 1 standard error of the mean. Total $N = 958$.



Supplementary Figure 11. Appropriateness ratings in response to a norm task as a function of condition and type of message (Democrat participants). Participants were asked to judge how appropriate it would be to post neutral vs. outrage messages to the social network they viewed depending on their condition. In the overperception condition, participants viewed messages in their newsfeed that tended to judge as containing more outrage than the author reported, as determined by Studies 1 and 2. In the accurate perception condition, participants viewed messages in their newsfeed that tended to be judged as containing similar outrage compared to the author's self-report. Box plots represent mean +/- 1 standard error of the mean. Total $N = 1066$.

Appendix A: The direct message sent to users in the Twitter field studies

Hi! I'm a researcher at Yale University, and my research group is interested in how people express themselves on social media. Would you like to answer a question to help us with our research? Your response will remain anonymous.

You sent the following tweet on [*date*]:

[*Tweet text displayed*]

Take a moment to think about what was happening at the time you tweeted. Think about who you were interacting with online, and what you were reading about on Twitter. Please answer the following regarding how you felt at the moment you posted the tweet:

1 How outraged did you feel on a 1-7 scale? (1 = not at all, 4 = somewhat, 7 = very)

2 How happy did you feel on a 1-7 scale? (1 = not at all, 4 = somewhat, 7 = very)

You can simply respond with one answer per line such as:

5

1