

Carriere, Kevin R.

Working Paper

A comment on "Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility"

I4R Discussion Paper Series, No. 309

Provided in Cooperation with:

The Institute for Replication (I4R)

Suggested Citation: Carriere, Kevin R. (2026) : A comment on "Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility", I4R Discussion Paper Series, No. 309, Institute for Replication (I4R), s.l.

This Version is available at:

<https://hdl.handle.net/10419/342979>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

No. 309
I4R DISCUSSION PAPER SERIES

**A comment on
“Overperception of moral outrage
in online social networks inflates
beliefs about intergroup hostility”**

Kevin R. Carriere

August 2026

I4R DISCUSSION PAPER SERIES

I4R DP No. 309

A comment on “Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility”

Kevin R. Carriere¹

¹Stonehill College, Easton/USA

AUGUST 2026

Any opinions in this paper are those of the author(s) and not those of the Institute for Replication (I4R). Research published in this series may include views on policy, but I4R takes no institutional policy positions.

I4R Discussion Papers are research papers of the Institute for Replication which are widely circulated to promote replications and meta-scientific work in the social sciences. Provided in cooperation with EconStor, a service of the [ZBW – Leibniz Information Centre for Economics](#), and [RWI – Leibniz Institute for Economic Research](#), I4R Discussion Papers are among others listed in RePEc (see IDEAS, EconPapers). Complete list of all I4R DPs - downloadable for free at the I4R website.

I4R Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Editors

Abel Brodeur

University of Ottawa

Jörg Ankel-Peters

RWI – Leibniz Institute for Economic Research

A comment on “Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility”

Kevin R. Carriere, Stonehill College
kcarriere@stonehill.edu

The author declares no conflict of interest.

The author declares no financial assistance for the creation of this manuscript.

Abstract

Brady et al. (2023) investigated whether people overperceive moral outrage in social media and whether exposure to outrage-saturated newsfeeds amplifies perceived collective outrage and intergroup hostility across five studies. Participants who viewed the high-overperception feed judged the collective outrage of their social network as substantially large. They also found that exposure to highly overperceived feeds increased perceived social appropriateness, increased affective polarization, and heightened perceived ideological extremity. In this commentary, I successfully computationally reproduce all reported effects even after correcting a data-processing error that unintentionally duplicated observations in Studies 1-3, including the large between-condition difference in perceived collective outrage (Study 4). Bootstrapping 2,000 resamples confirmed that the 95% bias-corrected and accelerated CI [2.09, 2.51] matches the analytic interval, indicating an exceptionally stable effect. I also replicate Study 5 using tweet-level mixed-effects models that preserve the nested structure of the data. These reanalyses reproduce the main effects of overperceived feeds on social appropriateness ($b = 0.67, p < .001$), ideological temperature ($\eta_p^2 = .27$), and ideological extremity ($b = 0.61, p < .001$). Finally, robustness checks excluding politically neutral participants, bootstrapping model estimates, and adding political ideology as a moderator all leave the magnitude and significance of the key findings unchanged, with the moderating effect of political ideology sparking additional future research questions. Overall, 100% of directional effects reproduce with near-identical size, demonstrating that the original paper's central claims about overperception-driven amplification of perceived outrage and polarization are computationally and statistically robust.

1. Introduction

The paper, Brady et al (2023), uses public, politically-focused Twitter posts from 2020 to investigate political partisans' perceptions of moral outrage in social media networks across five preregistered studies. Studies 1–3 estimated the degree of overperception of moral outrage among political partisans (identifying as Democrats or Republicans) Twitter users by comparing authors' self-reports to observers' ratings using multilevel models. Then, in Studies 4 and 5, they test whether exposure to a highly overperceived outrage-filled social media newsfeed causes overestimates of collective outrage (Study 4) and if it leads to further outrage expression, affective polarization, and perceived ideological extremity (Study 5) using experimental methods. The main results showed that individuals consistently overperceive others' moral outrage ($b \approx 0.6$, $p < .001$, 95% CI roughly [0.30, 0.90]) and that exposure to highly overperceived newsfeeds amplified perceived collective outrage and intergroup hostility, with large, standardized effects ($d \approx 0.40$ – 1.90).

In determining which claims I will computationally reproduce, I chose to focus on all hypothesis tests in the original study, while not focusing on reproducing tests reported in the supplementary material. In terms of sensitivity analyses, I had planned to test their robustness by utilizing robust standard errors and bootstrapping the mixed-effects models. Due to timing, the pre-registration did not get posted in time. Upon exploring the data in more detail, and after making the decision on the approach I had planned to take, it became clear that robust standard errors would not be more informative than the bootstrapped modeling that had been planned. As a result, I move forward with a subsampling stratification procedure as my main method of robustness checking but include other analyses that arose during the replication process, including parametric bootstrapping..

I am grateful to the original authors for providing the raw data and analysis scripts on the Open Science Framework. The pre-registrations can be found at Study 2 (<https://osf.io/ud5bc/overview>), Study 3 (<https://osf.io/qmnyb/overview>), Study 4 (<https://osf.io/sxtah/overview>), and Study 5 (<https://osf.io/mjftk/overview>). To obtain the data, I went to the Open Science Foundation page for where the manuscripts' data and scripts are stored (<https://osf.io/gtwsk/>) and downloaded them. The random seed was set to 131, and my full Markdown, along with support scripts for functions, is hosted at <https://osf.io/v3arj>. Studies 1 through 3 have a single script to reproduce the analyses, as their data cleaning methods and modeling procedures are similar. Users are instructed to simply replace the dataset reading for each variable. Studies 4 and 5 have their own scripts, which are also loaded into the Markdown file.

2. Computational Reproducibility

Scripts and data were downloaded from <https://osf.io/gtwsk/overview> and loaded into RStudio and combined into a single R Markdown file.

Across the five studies, I was able to reproduce all the target article's primary effects in terms of direction, statistical significance, and approximate effect size. Where code ran as written, the test statistics matched exactly or deviated only by rounding. Where code was missing or incorrect, I was able to reconstruct the analyses from the raw data, which was always the case for identifying confidence intervals of the results. At the effect level, the paper's central claims about overperception of outrage and the downstream consequences of outrage-saturated feeds are robust.

in contrast, workflow reproducibility and some specific inferential claims were weaker. Several analyses relied on undocumented derived datasets (of which, Study 4's could not be fully recreated), Study 4's stimulus file appears to contain at least one mis-labelled tweet, and several

descriptive statistics and confidence intervals are inaccurately reported. Code provided for the plots are not the plots provided in the manuscript.

Most importantly, when implementing the preregistered non-parametric analysis for Study 4, both high- and low-overperception feeds produce significant exaggeration of collective outrage, drawing some question into the paper's claim that only high-overperception networks lead to inflated collective judgments. In this sense, the effects reproduce, but one of the key boundary conditions does not when following the pre-registration.

Overall, I was able to reproduce the significance, direction, and effect sizes of all results throughout the manuscript. There are some places where code was not included within the scripts to replicate the analyses, but the correct analysis code was able to be written and confirmed.

Table 1.

Summary of computational reproducibility of Brady et al. (2023).

	Fully	Partial	No
Raw data provided	x		
Cleaning code provided	x		
Analysis data provided	x		
Analysis code provided		x	
Reproducible from raw data		x	
Reproducible from analysis data		x	

2.1 Discrepancies Between Pre-analysis Plan and Article

The authors pre-registered four out of five studies. The pre-registration plans for Studies 2, 3, and 4 note that paired t -tests at the tweet level would be used to compare conditions at the tweet-level, unless the data is non-normal. If the data was non-normal, they would use Wilcoxon-sign ranked tests.

However, the article tests its main predictions using multi-level models, treating the repeated ratings within participants and tweets as random effects and using the rating source (author vs. perceiver) as the fixed effect, as is pre-registered in Study 5¹. Although the text notes that a Wilcoxon signed-rank test was conducted for Study 3 (p. 925), no such analysis appears in the manuscript or in the accompanying scripts. Moreover, the analysis scripts for Studies 1-3 do not include any non-normality checks such as visually examining histograms or Q-Q plots, or statistical checks such as the Shapiro-Wilk test or the Kolmogorov-Smirnov test, and no note of departure from the pre-registrations for the analysis is noted.

It is worth noting, however, that the multi-level model is the better choice of analysis method than the paired *t*-test, and so the deviation is reasonable, but should have been reported (Willroth & Atherton, 2024). A paired *t*-test approach would have required collapsing each tweet into a single author–perceiver difference score, which loses the nested structure of the data. The multilevel model correctly accounts for the non-independence of ratings within tweets and within participants and provides a more appropriate test of overperception.

In contrast, Study 4’s script does include a visual inspection of the distributions of the variable per condition. While unreported, the script suggests that the authors deemed network outrage scores to have been normal and proceeded without using the non-parametric test in this case.

Studies 2 and 3’s pre-registration plans do not state that the authors will also collect confidence in one’s rating of outrage and happiness, but the variables appear within the dataset and are used in the supplemental material.

The paper reports a minimal number of analyses beyond the pre-registration, and these are noted within the paper to be exploratory in nature, notably the regressions in Studies 1 & 2

(combined) and Study 3 (by itself). Besides these deviations, the paper follows their pre-registrations as specified.

2.2. Errors that limited computational reproducibility

There was one significant coding error and a few minor coding and typographical errors that I uncovered while reproducing the manuscript. I also have some concern regarding an unaccounted for methodological choice of randomization, but I will address that later in my robustness checks.

The final datasets that are used to run the analyses of Studies 1 through 3 have an equal number of authors and perceiver observations. Since each perceiver rated almost thirty tweets and each author contributed only their self-rating, the ratio of perceiver-observations to author-observations should be roughly equal to the average number of ratings per perceiver. I identified that each author's self-rating was duplicated once for every perceiver who evaluated the same tweet, thereby increasing the number of observations in the dataset from roughly 4,800 to over 9,000. This occurred during a step in the data-cleaning process, where the authors attempted to merge information about the authors of the tweets with the observations of each respondent rating a given tweet. They merge their self-reported authors' observations by tweet id, which is correct, but do so on a dataset that is already set in a per-rating-Tweet level. Thus, for each rating of a single tweet, the author's observation attaches itself to the dataset. Instead, I add the authors on as an *additional* Tweet-level rating observation, and thereby, avoid the problem of having multiple single-author observations per Tweet.

First, I confirm that given a dataset set up in this way, the reported statistics in the manuscript reflect the computed results in the scripts. However, I also address this data-tidying in my own analysis and model a multi-level model that properly accounted for the nested data structure—ratings (level 1) made by multiple observers (level 2) on the same set of tweets (level 2 units). The

outcome variable was each observer's rating of a tweet's outrage (or happiness) on a scale of 1 to 7, and the key predictor was message source (author = self-report vs. perceiver = observer rating). Random intercepts were specified for both tweet and observer to account for non-independence across repeated ratings of the same tweet and repeated responses by the same observer. **Tables 2–7** show that the Conditional r^2 of the published models ranged from .753 to .851 but reduced to .519–.579 in my analyses. In these N -adjusted models, I also implemented a robustness check by estimating confidence intervals via profile-based methods (Pek & Wu, 2015). The third column reports 95% confidence intervals from a balanced partisan subsample and refit (1,000 iterations, equalizing Democratic and Republican perceivers per tweet), which are discussed in more detail in the robustness checks. Because the 2nd model reported changes the underlying observations, and the 3rd model reported is generated through a party-balanced subsampling-and-refit procedure, both models report their own coefficients for the variables under analysis. This approach is different than a parametric bootstrapping process, for which the reported fixed effect would not generally change, and the only differences would be seen within the standard error and the confidence intervals.

For the outrage models, the perceiver effect remains statistically significant in every study ($bs = 0.58$ – 0.63 , $ps < .001$), with overlapping 95% confidence intervals across the original and reanalyzed estimates, confirming that the main fixed-effect pattern is robust. In contrast, the happiness models remain nonsignificant throughout ($bs = -0.06$ to -0.17 , $ps > .27$). However, Cond r^2 values decline by roughly .25–.30 points in all studies (from .753–.851 to .519–.579), indicating that much of the initially reported variance explained at the random-effect level was attributable to the duplicated author rows. Marg r^2 values, which capture fixed-effect variance, remain nearly unchanged (.018–.022 vs. .002–.004), confirming that the replication preserves the substantive fixed effects but corrects inflated model-fit estimates arising from pseudo replication.

The corrected model yields a slightly higher precision around the fixed effect because the collinearity and overfitting issues associated with duplicated data have been resolved. Importantly, however, the overall results maintain their size, direction, and significance for the paper's overall findings. By adjusting the observations, of the six models reported in Studies 1 – 3, I find minor one-tenth of a point increase on overperception in outrage for Studies 2 ($b = 0.58$ to $b = 0.59$) and 3 ($b = 0.62$ to $b = 0.63$), and a one-tenth of a point decrease in the nonsignificant effect of overperception of happiness in Study 2 ($b = -0.17$ to $b = -0.16$) (See Tables 2 through 7).

Table 2.*Predicting Overestimated Outrage, Study 1*

	Computed Brady et al., 2023					Reanalysis – Fixed Obs.					Reanalysis – Fixed Obs. With Party-Stratified Subsample				
	b	SE	95% CI		p	b	SE	95% CI		p	b	SE	95% CI		p*
			LL	UL				LL	UL				LL	UL	
Intercept	3.955	0.175	3.611	4.30	<.001	3.955	0.167	3.628	4.282	<.001	3.959	0.152	3.663	4.247	<.001
Perceiver	0.591	0.231	0.139	1.044	.011	0.590	0.184	0.229	0.950	.002	0.586	0.151	0.291	0.895	<.001
Marg r^2			.018					.004							
Cond r^2			.847					.579							
PL-V			3.16					1.29							
TL-V			0.89					0.85							
R-V			0.75					1.57							

Note: Marg r^2 = Marginal r^2 ; Cond r^2 = Conditional r^2 ; PL-V = participant-level variance component; TL-V = tweet-level variance component; R-V = within-observation residual variance. p^* = Wald-style z p -value. All variance estimates are extracted from the random-effects structure of the multilevel model.

Table 3.
Predicting Overestimated Happiness, Study 1

	Computed Brady et al., 2023					Reanalysis – Fixed Obs.					Reanalysis – Fixed Obs. With Party-Stratified Subsample				
	<i>b</i>	<i>SE</i>	95% CI		<i>p</i>	<i>b</i>	<i>SE</i>	95% CI		<i>p</i>	<i>b</i>	<i>SE</i>	95% CI		<i>p</i> *
			<i>LL</i>	<i>UL</i>				<i>LL</i>	<i>UL</i>				<i>LL</i>	<i>UL</i>	
Intercept	3.128	0.158	2.818	3.437	<.001	3.128	0.162	2.811	3.445	<.001	3.138	0.145	2.852	3.422	<.001
Perceiver	-0.130	0.211	-0.544	0.284	.538	-0.130	0.191	-0.505	0.244	.495	-0.144	0.135	-0.419	0.123	.285
Marg r^2			.001					0							
Cond r^2			.839					.627							
PL-V			2.65					1.57							
TL-V			0.64					0.61							
R-V			0.63					1.30							

Note: Marg r^2 = Marginal r^2 ; Cond r^2 = Conditional r^2 ; PL-V = participant-level variance component; TL-V = tweet-level variance component; R-V = within-observation residual variance. p^* = Wald-style z p -value. All variance estimates are extracted from the random-effects structure of the multilevel model.

Table 4.
Predicting Overestimated Outrage, Study 2

	Computed Brady et al., 2023					Reanalysis – Fixed Obs.					Reanalysis – Fixed Obs. With Party-Stratified Subsample				
	<i>b</i>	<i>SE</i>	95% <i>CI</i>		<i>p</i>	<i>b</i>	<i>SE</i>	95% <i>CI</i>		<i>p</i>	<i>b</i>	<i>SE</i>	95% <i>CI</i>		<i>p</i> *
			<i>LL</i>	<i>UL</i>				<i>LL</i>	<i>UL</i>				<i>LL</i>	<i>UL</i>	
Intercept	4.010	0.145	3.726	4.293	<.001	4.005	0.140	3.730	4.280	<.001	4.024	0.136	3.761	4.283	<.001
Perceiver	0.584	0.171	0.249	0.919	.001	0.587	0.133	0.33	0.849	<.001	0.568	0.119	0.334	0.806	<.001
Marg r^2			.017					.003							
Cond r^2			.840					.581							
PL-V			2.58					0.81							
TL-V			1.47					1.43							
R-V			0.79					1.64							

Note: Marg r^2 = Marginal r^2 ; Cond r^2 = Conditional r^2 ; PL-V = participant-level variance component; TL-V = tweet-level variance component; R-V = within-observation residual variance. p^* = Wald-style z p -value. All variance estimates are extracted from the random-effects structure of the multilevel model.

Table 5.
Predicting Overestimated Happiness, Study 2

	Computed Brady et al., 2023					Reanalysis – Fixed Obs.					Reanalysis – Fixed Obs. With Party-Stratified Subsample				
	<i>b</i>	<i>SE</i>	95% <i>CI</i>		<i>p</i>	<i>b</i>	<i>SE</i>	95% <i>CI</i>		<i>p</i>	<i>b</i>	<i>SE</i>	95% <i>CI</i>		<i>p</i> *
			<i>LL</i>	<i>UL</i>				<i>LL</i>	<i>UL</i>				<i>LL</i>	<i>UL</i>	
Intercept	2.805	0.124	2.562	3.048	<.001	2.792	0.126	2.546	3.038	<.001	2.802	0.111	2.588	3.015	<.001
Perceiver	-0.170	0.162	-0.487	0.148	.295	-0.159	0.145	-0.443	0.125	.274	-0.168	0.104	-0.367	0.041	.105
Marg r^2			.002					0							
Cond r^2			.843					.633							
PL-V			2.32					1.34							
TL-V			0.66					0.63							
R-V			0.56					1.14							

Note: Marg r^2 = Marginal r^2 ; Cond r^2 = Conditional r^2 ; PL-V = participant-level variance component; TL-V = tweet-level variance component; R-V = within-observation residual variance. p^* = Wald-style z p -value. All variance estimates are extracted from the random-effects structure of the multilevel model.

Table 6.*Predicting Overestimated Outrage, Study 3*

	Computed Brady et al., 2023					Reanalysis – Fixed Obs.					Reanalysis – Fixed Obs. With Party-Stratified Subsample				
	<i>b</i>	<i>SE</i>	95% <i>CI</i>		<i>p</i>	<i>b</i>	<i>SE</i>	95% <i>CI</i>		<i>p</i>	<i>b</i>	<i>SE</i>	95% <i>CI</i>		<i>p</i> *
			<i>LL</i>	<i>UL</i>				<i>LL</i>	<i>UL</i>				<i>LL</i>	<i>UL</i>	
Intercept	4.005	0.131	3.748	4.263	<.001	4.005	0.135	3.740	4.270	<.001	4.052	0.130	3.783	4.299	<.001
Perceiver	0.623	0.129	0.370	0.876	<.001	0.629	0.115	0.404	0.854	<.001	0.577	0.118	0.348	0.809	<.001
Marg r^2			.002					0							
Cond r^2			.843					.633							
PL-V			2.32					1.34							
TL-V			0.66					0.63							
R-V			0.56					1.14							

Note: Marg r^2 = Marginal r^2 ; Cond r^2 = Conditional r^2 ; PL-V = participant-level variance component; TL-V = tweet-level variance component; R-V = within-observation residual variance. * = Wald-style z p -value. All variance estimates are extracted from the random-effects structure of the multilevel model.

Table 7.
Predicting Overestimated Happiness, Study 3

	Computed Brady et al., 2023					Reanalysis – Fixed Obs.					Reanalysis – Fixed Obs. With Party-Stratified Subsample				
	<i>b</i>	<i>SE</i>	95% <i>CI</i>		<i>p</i>	<i>b</i>	<i>SE</i>	95% <i>CI</i>		<i>p</i>	<i>b</i>	<i>SE</i>	95% <i>CI</i>		<i>p</i> *
			<i>LL</i>	<i>UL</i>				<i>LL</i>	<i>UL</i>				<i>LL</i>	<i>UL</i>	
Intercept	2.792	0.113	2.571	3.013	<.001	2.792	0.122	2.553	3.030	<.001	2.798	0.110	2.580	3.014	<.001
Perceiver	-0.057	0.127	-0.305	0.192	.656	-0.055	0.124	-0.299	0.188	.657	-0.061	0.104	-0.261	0.141	.558
Marg <i>r</i> ²	.002					0									
Cond <i>r</i> ²	.843					.633									
PL-V	2.32					1.34									
TL-V	0.66					0.63									
R-V	0.56					1.14									

Note: Marg r^2 = Marginal r^2 ; Cond r^2 = Conditional r^2 ; PL-V = participant-level variance component; TL-V = tweet-level variance component; R-V = within-observation residual variance. p^* = Wald-style z p -value. All variance estimates are extracted from the random-effects structure of the multilevel model.

2.3 Code provides no confidence interval output and limited post-hoc testing

For all studies, the code necessary to compute confidence intervals is missing from the analysis scripts for their multi-level models. I address this in all analyses by generating profile-based confidence intervals (Pek & Wu, 2015), which matches the size, direction, and significance of the reported results. There are minor deviations from the reported confidence intervals in the paper from the computational replication (See **Table 8**, where Study 3's two main results differ slightly, and Study 4's main result is most likely a rounding discrepancy (the same with Study 2 on Happiness). These differences are minor and do not change the direction, magnitude, or substantive interpretation of the results. The discrepancies likely reflect the use of a different estimation method for the confidence intervals, but without having the method used by the authors, it is not clear where this discrepancy lies.

A similar difficulty arose in Study 5, where participants were asked to judge how their presented social network felt about the political outgroup and ingroup using a feeling thermometer. A mixed-effects ANOVA predicting response on the feeling thermometer through the between-subjects effect of social network feed condition and the within-subjects effect of the network being responded about (ingroup or outgroup) was conducted. While I can computationally reproduce the reported interaction effect, $F(1,1010) = 369.58, p < .001$, the scripts do not conduct any post-hoc tests or simple main effect tests, and thus, I am unable to replicate the post-hoc comparisons by the material alone.

To reproduce the analysis, I computed simple main effects of condition within network using Welch two-sample t-tests using the *rstatix* package. For the ingroup network, perceived temperature was higher in the high-overperception condition than in the Low-Overperception condition ($M = 75.07, SD = 21.14, M = 56.12, SD = 24.96, t(958.97) = 12.99, p < .001, d = 0.82$). For the outgroup network, perceived temperature was lower in the High- than Low-Overperception condition, $M_{\text{High}} =$

15.64, $SD_{\text{High}} = 20.26$, $M_{\text{Low}} = 48.23$, $SD_{\text{Low}} = 24.53$, $t(948.38) = -22.95$, $p < .001$, $d = -1.44$. These results fully reproduce the reported pattern: High-Overperception increases perceived warmth toward ingroup content and decreases it toward outgroup content, with very large directionally consistent effects.

Table 8.

Computational comparisons of the confidence intervals

	Reported 95% CI		Computed 95% CI		Replicated CIs
	LL	UL	LL	UL	
Study 1 Outrage	0.14	1.04	0.14	1.04	x
Study 1 Happiness	-0.54	0.28	-0.54	0.28	x
Study 2 Outrage	0.25	0.92	0.25	0.92	x
Study 2 Happiness	-0.49	0.14	-0.487	0.148	---
Study 1 & 2 Correlation	0.06	0.31	0.06	0.31	x
Study 3 Outrage	0.27	0.93	0.37	0.88	---
Study 3 Happiness	-0.50	0.11	-0.30	0.19	---
Study 3 Correlation	0.08	0.32	0.08	0.32	X
Study 4 Mean Diff.	2.09	2.50	2.088	2.506	---
Study 4 One-Sample T-Test High.	5.69	5.95	5.694	5.951	x
Study 4 One Sample T-Test Low	3.36	3.69	3.360	3.692	x
Study 5 Social Appropriateness	0.47	0.85	0.47	0.85	x
Study 5 Ideological Extremity	0.46	0.64	0.46	0.64	x

2.4 Recreating the dataset from Study 4

As noted earlier, the analysis scripts load several derived datasets that were created from the original data files, but the scripts do not include the code used to generate these intermediate datasets. This undocumented preprocessing makes it difficult to trace variable construction or ensure data-key reliability across studies. Although the derived datasets used in Studies 1 through 3

can be reconstructed with reasonable accuracy from the raw data, the same is not true for Study 4's preprocessed dataset. The Study 4 stimulus file appears to contain at least one mis-transcribed stimulus entry.

Specifically, the dataset includes a stimulus labeled tweet #90, but no tweet with ID 90 in any raw dataset matches the outrage statistics attributed to this stimulus. The derived dataset lists its perceived outrage as $M = 5.5769$ with an author-reported outrage score of 1.00. However, the corresponding means for tweet 90 across the raw datasets are $M_{\text{Study1}} = 5.00$, $M_{\text{Study2}} = 5.0377$, and $M_{\text{Study3}} = 4.7955$, none of which align with the derived value.

Because each tweet has many participant ratings, duplicate point estimates are extremely rare. This allows us to reverse-identify the likely source stimulus by matching both the perceived-outrage mean and the author's self-reported score. Only two tweets in the raw data share the combination $M = 5.5769$ and author outrage = 1.00: tweet 85 and tweet 89, both from Study 2. Tweet 85 already appears correctly in the derived Study 4 dataset, leaving tweet 89 as the only viable source of the mislabeled stimulus. Therefore, the entry labeled "#90" in the Study 4 file is almost certainly intended to be tweet #89.

Still, the disconnects throughout the datasets made it difficult to confirm that the author's political ideologies from the original tweets were consistent with the provided data, as the provided dataset for author information from Study 2 does not include the author's ideology scores, unlike Study 1.

2.4 Study 5 includes misexposed participants

In Study 4, Brady et al. (2023) note that "in both conditions, participants only viewed political ingroup tweets based on their self-reported party identification (any participants who did not identify as Republican or Democrat ($N = 18$) were removed from the analysis)" (p. 925)". They repeat the same procedure in Study 5. However, when tabulating the self-reported party identification against the

experimental conditions, there are 40 observations (4 participants at 10 observations per participant) that identified as Republican but were exposed to Democrat social media feeds (IDs 430, 517, 920, and 983) in Study 5. To verify that these participants were indeed politically distinct from the Democratic group, I conducted a paired-samples t-test comparing their feeling thermometer ratings toward Republicans ($M = 69.74$, $SD = 14.06$) against Democrats ($M = 26.50$, $SD = 19.96$), $M_{\text{Diff}} = -43.25$, $SD_{\text{Diff}} = 23.25$, $t(3) = -2.63$, $p = .078$, $d = 1.31$. While this difference was not statistically significant, the effect size is large for a small sample size.

Using their mean as the comparison mean against identified Democrats, I also ran a one-sample t-test to test if Democrat-identified participants felt differently towards the claimed political ingroup (Democrats) and the claimed political outgroup (Republicans). Democrat participants felt warmer to Democrats ($M = 73.80$, $SD = 18.36$) than the four's average reported thermometer, $t(531) = 59.28$, $p < .001$, $d = 2.57$, 95% CI [72.13, 75.26], and also felt cooler towards Republicans ($M = 14.420$, $SD = 16.10$) than the four's average reported thermometer to Republicans, $t(531) = -79.26$, $p < .001$, $d = 3.44$, 95% CI [13.04, 15.79]. Based on this analysis, I conclude that these four participants are more likely than not to be conservative but were exposed to Democratic news profiles. To ensure that the data remains focused on individuals responding to profiles of their political ingroup, I proceeded with testing the robustness of Study 5 by excluding the data from these four participants.

2.5 Minor reporting errors

There are a few minor reporting errors within the manuscript when it comes to means. First, I find that there were 165 observers in Study 2, not 189. Study 5 reports 1,013 as the final N after exclusion criteria, but within that 1,013 is an observation with no data (only missing values), that is unknowingly filtered out during data cleaning, and I find reporting the final N for Study 5 to more accurately be 1,012, as that is the number of observations within the dataset when the analysis of

any data begins. Study 5 describes a Bonferroni-corrected d adjustment instead of a Bonferroni-corrected α level. There are a few instances where the sign of the test statistic is reversed from what it should be (Study 5, social appropriateness and Study 5, ideological extremity), but the confidence intervals and effect sizes are in the correct direction, indicating that it was likely a typographical error.

Study 4 also contains a number of inconsistencies in the reporting of point estimates across its measures. The first set of discrepancies concerns the overperception averages used to describe the tweet feeds. All four reported values deviate from the data, and the magnitude of overperception is understated, with the accurate values further emphasizing the differences in conditions. They report an overperception average of $M_{Republican} = 4.25$ and $M_{Democrat} = 3.92$ for the high-overperception condition and $M_{Republican} = 1.58$ and $M_{Democrat} = 1.36$ for the low-overperception conditions but I find that the data shows an overperception average of $M_{Republican} = 4.27$ and $M_{Democrat} = 3.93$ for the high-overperception condition and $M_{Republican} = 0.75$ and $M_{Democrat} = 1.37$.

A second inconsistency appears in the description of the one-sample test comparing collective outrage judgments against the average outrage levels of the tweets in each feed. The paper reports that the mean collective outrage judgment in the high-overperception condition was $M = 5.96$ and that this was compared against 4.09, the mean of the individually perceived outrage values for the high overperception tweets. In contrast, their data and scripts, and my own analysis, finds that the collective outrage mean is $M = 5.82$ and that number was compared against $M = 4.94$.

Across the five studies, my computational reproduction largely reproduced the key findings reported by Brady et al. (2023), though some discrepancies emerged. I verified most descriptive statistics, effect directions, and inferential outcomes, including the consistent overperception of moral outrage across Studies 1–3, the network-level amplification of outrage in Study 4 (with an additional significant median difference in perceived collective outrage for low overperception), and

the effects of overperceived feeds on social appropriateness and ideological extremity in Study 5. Some of the limits of my computational reproducibility was due to a lack of code, while others appear to be nothing more than a rounding difference. Despite these discrepancies, the reproduced estimates closely match the reported magnitudes, supporting the robustness of the central claims while revealing minor reporting and documentation issues in the original analysis workflow.

3. Robustness Reproduction

I now turn my attention to testing the robustness of their findings. I test the robustness of the effects using subsampling and refitting methods on my multi-level models. The decision to conduct these tests was taken after reading the paper but prior to observing the scripts or data. However, in the process of examining the data and exploring its computational reproducibility, I had further questions regarding the data and how addressing the modeling and analytic procedures around the data would influence the results. I detail all questions and answers below.

3.1 Party-Stratified Subsampling Within Tweet Sets

In Brady et al. (2023), the authors note that “each participant judged a stratified random selection of 30 tweets (15 were randomly drawn from those classified by DOC as outrage and 15 were randomly drawn from those classified by DOC as non-outrage)” (p. 925). However, this goal was not fully achieved in any of the studies, with only 44.57% of participants across the first three studies rating a total of 30 tweets. Although a substantial proportion of participants did not rate 30 total tweets, a more consequential issue emerges when examining how the tweets were distributed (see **Table 9**).

Table 9.

Distribution of tweets and percent rated out of 30 for Studies 1 through 3

Percent of Tweets Rated	Study 1 _{Percent}	Study 2 _{Percent}	Study 3 _{Percent}
0.00%			1.10%

6.67%			0.60%
13.33%			0.60%
23.33%			0.60%
50.00%			0.30%
56.67%			0.30%
63.33%	21.40%	0.5%	0.60%
66.67%		13.2%	13.00%
80.00%			0.60%
90.00%	19.30%		
93.33%	17.90%	29.60%	13.80%
96.67%	21.40%	14.30%	13.50%
100.00%	20%	42.30%	55.20%

Rather than being randomly sampled from the full corpus (as implied by the term stratified random selection), participants were each assigned to one of a few fixed tweet sets (See **Tables 10-12**). Within a set, all participants rated the same tweets, and the number of tweets per set (and ratings per participant) was constant. Consequently, the number of tweets rated per participant and the number of ratings per tweet were perfectly confounded with tweet set membership. In other words, exposure was not only unequal across participants but also structurally clustered by tweet set.

Table 10.

Tweet ID Groups and Ideological Distributions for Study 1

	Total # of Participants	# of Democrats	# of Republicans
61,62,63,65,66,67,68,69,70,133,134,135,136, 137,138,139,140,141,142	24	13	11
16,17,18,19,20,21,22,23,24,25,26,27,28,29,30, 88,89,90,91,92,93,94,95,96,97,98,99,100,101,102	23	11	12

46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,118, 119,120,121,122,123,124,125,126,127, 128,130,131,132	22	11	11
1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,73,74,75, 76,77,78,79,80,81,85,86,87	21	10	11
31,32,33,34,35,36,37,38,40,41,42,43,44,45, 103,104,105,106,107,108,109,110,111,113,114, 115,116,117	20	8	12

Table 11.*Tweet ID Groups and Ideological Distributions for Study 2*

	Total # of Participants	# of Democrats	# of Republicans
76,77,78,79,80,82,83,84,85,86,87,89,90,176, 177,178,179,180,181,182,183,184,185,186, 187,188,189,190	26	14	12
1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,101,102, 103,104,105,106,107,108,109,110,111,112, 113,114,115	25	12	13
31,32,33,34,35,36,37,38,41,42,43,44,45,131, 132,133,134,135,136,137,138,139,140,141, 142,143,144,145	25	10	15
46,47,48,49,50,51,52,53,54,55,56,57,58,59, 60,146,147,148,149,150,151,152,153,154, 155,156,157,158,159,160	25	12	13
61,62,63,65,66,67,68,69,70,71,72,73,74,75, 161,162,163,164,165,166,167,168,169,170, 171,172,173,174,175	25	14	11
16,17,18,19,20,21,22,23,24,25,26,27,28,29, 30,116,117,118,119,120,121,122,123,124, 125,126,127,128,129,130	22	10	12
91,92,93,94,95,96,97,98,99,100,191,192,193, 194,195,196,197,198,199,200	17*	8*	9

Table 12.*Tweet ID Groups and Ideological Distributions for Study 3*

	Total # of Participants	# of Democrats	# of Republicans
--	----------------------------	-------------------	---------------------

16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,116, 117,118,119,120,121,122,123,124,125,126,127,12 8,129,130	49	27	22
1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,101,102, 103,104,105,106,107,108,109,110,111,112, 113,114,115	46	22	24
61,62,63,65,66,67,68,69,70,71,72,73,74,75, 161,162,163,164,165,166,167,168,169,170, 171, 172,173,174,175	44	24	20
76,77,78,79,80,81,82,83,84,85,86,87,88,89, 90,176,177,178,179,180,181,182,183,184,185, 186,187,188,189,190	44	23	21
91,92,93,94,95,96,97,98,99,100,191,192,193, 194,195,196,197,198,199,200	45**	21**	24
31,32,33,34,35,36,37,38,41,42,43,44,45,131, 132,133,134,135,136,137,138,139,140,141, 142,143,144,145	41	21	20
46,47,48,49,50,51,52,53,54,55,56,57,58,59, 60,146,147,148,149,150,151,152,153,154, 155,156,157,158,159,160	41	23	18

** Each * indicates a participant that did not fully complete their sets and are marked as different sets in my code. However, they are added to the appropriate set they would have been placed in, should they have completed the study in full.*

I deviated from my pre-registration to investigate whether any set of tweets was judged as more outraged than others. To do so, I ran a multilevel model with tweet and participant as random effects and tweet set as the between effect, with separate models run to compare outrage-based judgments and happiness-based judgments. I did this once for each study. Analyzing these model through an ANOVA framework, I find no significant effect on tweet set on ratings of outrage, $F_{\text{Study1}}(4.00, 212.92) = 0.20, p = .940$, $F_{\text{Study2}}(6.00, 303.35) = .28, p = .945$, or $F_{\text{Study3}}(6.00, 258.14) = 0.22, p = .970$ nor on happiness, $F_{\text{Study1}}(4.00, 177.69) = 0.24, p = .916$, $F_{\text{Study2}}(6.00, 288.43) = 0.33, p = .921$, or $F_{\text{Study3}}(6.00, 441.81) = 0.78, p = .587$. There also was no difference in the signed difference score based on tweet set for outrage, $F_{\text{Study1}}(4.00, 165.06) = 1.05, p = .384$, $F_{\text{Study2}}(6.00, 240.09) = 0.99, p = .435$, $F_{\text{Study3}}(6.00, 212.54) = 1.35, p = .236$ nor happiness, $F_{\text{Study1}}(4.00, 204.10) = 0.23, p = .922$, $F_{\text{Study2}}(6.00, 293.48) = 0.83, p = .545$, $F_{\text{Study3}}(6.00, 250.13) = 0.68, p = .666$.

Even though they did not show any significant effects on their own, for completeness, I modeled the original multi-level models predicting perceptions of outrage by perspective (author v. observer) and tweet-set as an additional between-subjects control and included tweet and participant as the random effects. **Table 13** reports the coefficient for the indicator variable of whether the judgment of emotion came from the author or the perceiver. The *p*-values for all coefficients of interest do not shift, but the point estimates do move slightly.

Table 13.

Differences in key coefficients In Studies 1 through 3 controlling for tweet-set

	Brady et al. (2023) Estimates				Tweet-Set Controlled			
	<i>b</i>	<i>SE</i>	95% CI		<i>b</i>	<i>SE</i>	95% CI	
			LL	UL			LL	UL
Study 1 Outrage	0.591	0.231	[0.139, 1.044]	.011	0.604	0.186	[0.242, 0.966]	.001
Study 1 Happiness	-0.130	0.211	[-0.544, 0.284]	.538	-0.139	0.192	[-0.514, 0.235]	.469
Study 2 Outrage	0.584	0.171	[0.249, 0.919]	.001	0.580	0.134	[0.318, 0.841]	<.001
Study 2 Happiness	-0.170	0.162	[-0.487, 0.148]	.295	-0.171	0.148	[-0.459, 0.116]	.247
Study 3 Outrage	0.623	0.129	[0.370, 0.876]	<.001	0.572	0.110	[0.356, 0.788]	<.001
Study 3 Happiness	-0.057	0.127	[-0.305, 0.192]	.656	-0.061	0.123	[-0.301, 0.180]	.621

Note: Brady et al. (2023) was already reported (See Tables 2-7); they are provided for comparability purposes only.

In addition, the paper had noted that the goal of this randomization was “to have each tweet rated by a minimum of ten Democrats and ten Republicans” (p. 924). Yet, as seen in **Table 10**, this was not always the case. I wanted to know whether accounting for this inequality of partisanship across tweets would impact judgments of moral emotions. As a result, I implemented a 1,000-iteration subsample that sampled equal numbers of Democratic and Republican perceivers per tweet (8 per group) and refit the model each time. I chose eight as the sample size because it was the minimum of a given political party within the datasets (See **Tables 10 & 11**). Despite this, the

mean coefficient and confidence interval remained stable, confirming that overperception does not depend on tweet-level clustering. Across all three studies, the mean coefficients and percentile confidence intervals were nearly identical to the fixed-effects estimates (see **Table 2** through **Table 7**). This indicates that the observed overperception effects are robust to tweet-level clustering and partisan imbalance in exposure. In other words, participants' tendency to overestimate outrage (but not happiness) was consistent even when equalizing tweet-by-party sampling density.

3.2 Parametric Bootstrapping of Multi-Level Models

While I have already discussed the balanced, tweet-level subsample used to equalize partisan exposure, that procedure was not preregistered. Instead, I had planned a basic bootstrap multi-level model. Using 1,000 resamples of the fitted models for each study, I re-estimated the random-effects structure and recorded the fixed effect of perspective (author vs. perceiver) on perceived outrage and happiness. This approach simulates new outcomes from the fitted model to generate distributions of the fixed-effects estimates, allowing for confidence intervals without relying on normal-approximation assumptions (See **Table 14**). The results show no substantive differences from what was already discussed, and so we move on from this analysis quickly, and provide it purely for completeness.

Table 14.*Pre-Registered Bootstrapping Analysis of Studies 1 through 3*

	Study 1: Outrage						Study 1: Happiness					
	$\hat{\beta}$	$\bar{\beta}$	SE	95% CI		p	$\hat{\beta}$	$\bar{\beta}$	SE	95% CI		p
				LL	UL					LL	UL	
Intercept	3.955	3.961	0.162	3.646	4.277	<.001	3.128	3.133	0.158	2.822	3.431	<.001
Perceiver	0.589	0.587	0.184	0.222	0.950	.001	-0.130	-0.132	0.191	-0.514	0.245	.495
	Study 2: Outrage						Study 2: Happiness					
	$\hat{\beta}$	$\bar{\beta}$	SE	95% CI		p	$\hat{\beta}$	$\bar{\beta}$	SE	95% CI		p
				LL	UL					LL	UL	
Intercept	4.017	4.020	0.144	3.739	4.299	<.001	2.792	2.792	0.122	2.546	3.028	<.001
Perceiver	0.576	0.572	0.139	0.302	0.849	<.001	-0.159	-0.160	0.146	-0.451	0.105	.275
	Study 3: Outrage						Study 3: Happiness					
	$\hat{\beta}$	$\bar{\beta}$	SE	95% CI		p	$\hat{\beta}$	$\bar{\beta}$	SE	95% CI		p
				LL	UL					LL	UL	
Intercept	4.059	4.060	0.127	3.817	4.317	<.001	2.795	2.796	0.116	2.574	3.033	<.001
Perceiver	0.572	0.572	0.108	0.368	0.783	<.001	-0.059	-0.060	0.121	-0.283	0.178	.625

Study 4 used an independent samples *t*-test to examine whether perceived collective outrage differed by overperception newsfeed conditions. I implemented a non-parametric bootstrap of the mean difference (2,000 resamples). This yielded a 95% BCa confidence interval of [2.10, 2.51], virtually identical to the analytic *t*-test interval of $t(479.50) = 21.56, p < .001, d = 1.91, 95\% CI [2.09, 2.51]$, confirming the robustness of this large effect.

3.3 Non-Parametric Tests in Study 4

As noted in 2.1, the authors preregistered to use nonparametric tests should the data be non-normal, and in the script for Study 4, the authors did visually inspect the data before proceeding with an independent samples *t*-test. To further pressure-test this result, I used a Shapiro-Wilk test to examine the distribution of the data. Because the distributions of perceived outrage were significantly non-normal in both conditions (Shapiro-Wilk tests: $W = 0.94, p < .001$, for low-overperception; $W = 0.85, p < .001$, for high-overperception), nonparametric Wilcoxon signed-rank tests were used to compare participants' median collective outrage ratings to the median perceived outrage across their feeds. Supporting the weighted-average prediction, the results showed that in the high-overperception condition, participants' perceptions of collective outrage (Median = 6.00) were significantly greater than the median perceived outrage across their individual feed messages (Median = 5.50), $W = 24,299.5, p < .001, r = .34$. Interestingly, in the low-overperception condition, participants' perceptions of collective outrage (Median = 4.00) were also significantly greater than the average perceived outrage across their messages (Median = 3.00), $W = 12,810, p < .001, r = .35$, going against the published findings, which only reported a significant effect in the high-overperception condition. Under the preregistered nonparametric approach, both conditions show that participants systematically judged the collective to be more outraged than the underlying tweet content would warrant.

3.4 Preserving the data-structure of Study 5's acceptability measure

In the original analysis of Study 5, Brady et al. (2023) used an independent-samples *t*-test to compare whether participants exposed to a high-overperception newsfeed judged novel tweets as more socially acceptable than those in a neutral condition. This analytic approach required averaging across the four to five outrage tweets and the five neutral tweets per participant—collapsing 9,644 individual observations into 1,012 participant-level means, a 9.5x reduction in statistical power. Converting the reported *t*-statistic (6.89; $d = 0.43$) yields an approximate R^2 of .05, suggesting a small overall effect.

For any *t*-test, I pre-registered to use a nonparametric test to assess the result. Running a Mann-Whitney *U*-test on the condition for average appropriateness (averaged at the participant-level, excluding the four misexposed participants) was significant, $W = 154702$, $p < .001$, $r = -.190$. I deviate from the pre-registered robustness analyses to further examine this modeling choice.

To preserve the full tweet-level variability and account for the nested structure of the data, I reanalyzed the study using a linear mixed-effects model with random intercepts for both participant and tweet. Unlike the *t*-test, which assumes independence among averaged observations, the hierarchical model treats tweets as repeated measures nested within participants—allowing tweet-level variance and participant-level variance to be modeled simultaneously and test whether the social acceptability of outrage tweets depends on exposure to outrage-saturated newsfeeds. Beyond that, I can examine whether the effect of the network condition is moderated by emotion expressed in judging social acceptability.

The model explained a modest proportion of variance ($R^2 = .12$), consistent with the original but now estimated from the full 9,644 tweet-level ratings rather than participant averages. The interaction between condition and tweet type was large and significant, $b = 0.667$, $SE = 0.055$,

$t(8,620.002) = 12.170, p < .001, 95\% \text{ CI } [0.559, 0.774]$, indicating the effect of exposure depends on whether the tweet expressed outrage. In the Low-overperception condition, outrage tweets were judged far less socially appropriate ($M = 0.13, SE = 0.20$) than neutral tweets ($M = 1.65, SE = 0.19$), $p = .002$. In the High-overperception condition, however, this difference was attenuated and nonsignificant ($M_{\text{High, Outrage}} = 0.89$ vs. $M_{\text{High, neutral}} = 1.75$), $p = .410$. Pairwise contrasts confirmed that exposure to high-overperception feeds significantly increased perceived appropriateness of outrage tweets ($p < .001$) but did not significantly change ratings of neutral tweets ($p = .500$).

3.5 - Affective-polarization analysis using residual correlations

The authors report minimal statistics on the analyses of their feeling thermometer outcome in Study 5, but they do describe a clear condition x target pattern in participants' judgments of how their social network felt about the political ingroup and outgroup. Specifically, participants in the high-overperception condition judged their social network to like the ingroup more ($M = 75.07$ vs. 56.12) and to dislike the outgroup more ($M = 15.64$ vs. 48.23) compared to those in the low-overperception condition. Post-hoc comparisons further indicated that the effect of condition on outgroup dislike ($M_{\text{diff}} = 32.59$) was nearly twice as large as the effect on ingroup liking ($M_{\text{diff}} = 18.95$), leading the authors to conclude that exposure to high-overperception newsfeeds substantially increased perceived affective polarization.

To account for the repeated-measures nature of the thermometer ratings (participants rated thermometers of both their own and the opposing party), we fitted a linear mixed-effects model using the nlme package in R. This specification included a random intercept for participant and a compound-symmetry correlation structure to model within-participant covariance between the two ratings. This approach allows the model to account for non-independence of repeated observations without aggregating across targets. Modeling with this correlational structure provided a better fit to

the data ($\Delta AIC = \sim 800$) and appropriately modeled within-person covariance, while a non-specified correlational structure model with random intercepts converged to a singular fit (participant-level variance = 0).

Results were directionally consistent with Brady et al.'s (2023) description: the condition $\times$ target interaction was significant, indicating a larger own–other party warmth gap in the high-overperception condition, $b = 51.977$, $SE = 2.664$, $t(1006) = 19.508$, $p < .001$. The main effects of condition $b = -32.808$, $SE = 1.424$, $t(1006) = -23.040$, $p < .001$, and target $b = 8.031$, $SE = 1.914$, $t(1006) = 4.196$, $p < .001$, were also significant. Simple-effects contrasts showed greater ingroup warmth under high versus low overperception, $\Delta M = 19.17$, $SE = 1.42$, $t(1006) = 13.46$, $p < .001$, and markedly lower outgroup warmth under high versus low overperception, $\Delta M = 32.81$, $SE = 1.42$, $t(1006) = 23.04$, $p < .001$. Within-condition ingroup–outgroup gaps were 60.01 ($SE = 1.85$), $t(1006) = 32.37$, $p < .001$, in the high-overperception condition and 8.03 ($SE = 1.91$), $t(1006) = 4.20$, $p < .001$, in the low-overperception condition. Substantively, these model-based EMMs support the same qualitative inference as the original article: exposure to high-overperception newsfeeds is associated with greater perceived affective polarization, driven primarily by colder outgroup judgments and warmer ingroup judgments.

3.6 – Increasing the partisanship of Study 5's sample

The manuscript excludes individuals who had not identified as Democrat or Republican as a representation of partisanship. However, upon examining the data, 18 participants have equal feeling thermometers for both political parties, and a total of 43 participants have thermometers that differ by less than 5 (including zero). As a robustness check, I exclude these 43 participants to examine if a more restrictive sample changes the substantive significance of the findings. I find, and

report below, that such changes have no visible impact on the size, significance, or direction of the findings.

The original effect size for social appropriateness was $d = 0.43$ with a mean difference of $M_{\text{Diff}} = 0.66$ points on a scale of 1 to 7. Reducing the sample to ensure a more polarized sample population, I still replicate the significant effect of the overperception network, $t(936.26) = 6.42, p < .001, d = 0.41$. Individuals who were exposed to high overperceived outrage networks ($M = 1.34, SD = 0.89$) were more likely to rate highly outraged tweets and morally acceptable compared to low outraged feeds ($M = 0.93, SD = 1.11$). The second result of affective polarization was reported by describing the interaction, which the computational analysis showed had an $\eta_p^2 = .268$. By reducing the sample size to a more partisan sample, I find an increase in the effect size of the interaction to $\eta_p^2 = .286$ (See Table 15 for the means), and the mixed effects model showed almost no movement in the estimate of the interaction, $b = 51.988, SE = 2.725, t(963) = 19.082, p < .001$. However, the reduced sample showed no change in comparison of perceived ideological extremity of one's newsfeed, $t(960) = 11.31, p < .001, d = .72$, with the effect remaining the same size as without it.

Table 15.

Means and standard deviations of Study 5's affective polarization results

	Published Results		Higher polarization sample	
	High-Overperception	Low-Overperception	High-Overperception	Low-Overperception
Ingroup Network	75.07 (21.14)	56.12 (24.96)	75.33 (20.87)	56.35 (24.79)
Outgroup Network	15.64 (20.26)	48.23 (24.53)	15.37 (20.12)	48.38 (24.42)

3.7 Adding political ideology as a moderating variable

Finally, I turn to examining the robustness of their findings when controlling for political ideology of the participant. Overall, the effects remain robust to these controls, but the analyses sheds some additional light from examining the moderating effect of political ideology.

As shown in **Table 16**, results were consistent across the first three studies: observers overperceived outrage more than happiness overall, Republicans tended to show larger signed deviations than Democrats, and the Republican–Democrat difference was attenuated when judging outrage specifically.

Table 16.*Predicting Differences in Judgments by Emotional Focus and Political Party, Studies 1-3*

	Study 1					Study 2					Study 3				
	<i>b</i>	<i>SE</i>	95% <i>CI</i>		<i>p</i>	<i>b</i>	<i>SE</i>	95% <i>CI</i>		<i>p</i>	<i>b</i>	<i>SE</i>	95% <i>CI</i>		<i>p</i>
			<i>LL</i>	<i>UL</i>				<i>LL</i>	<i>UL</i>				<i>LL</i>	<i>UL</i>	
Intercept	-0.50	0.16	-0.81	-0.19	.002	-0.40	0.12	-0.63	-0.17	.001	-0.16	0.10	-0.36	0.04	.110
Outrage	1.13	0.08	0.97	1.29	<.001	1.02	0.06	0.90	1.14	<.001	0.78	0.04	0.69	0.86	<.001
Republican	0.69	0.15	0.40	0.99	<.001	0.44	0.11	0.23	0.65	<.001	0.24	0.08	0.09	0.39	.001
Outrage* Republican	-0.75	0.11	-0.97	-0.53	<.001	-0.50	0.09	-0.67	-0.34	<.001	-0.22	0.06	-0.34	-0.09	.001
Marg r^2			.03					.03					.02		
Cond r^2			.34					.32					.31		

The last two studies created faux networks based on the tweets authored by Democrat and Republican tweet authors. These networks were constructed in such a way where one network included ten tweets that were generally highly overperceived in their outrage, while the other network was low in its overperception. The feeds themselves were designed in such a way that they would be of the political in-group of the participant.

Study 4 tested whether participants judged their assigned networks (high vs. low overperception) as more outraged when viewed collectively compared to how the same tweets were judged individually in Studies 1–3. The authors reported that only the high-overperception network produced collective outrage estimates above the individual tweet averages. However, using the nonparametric approach the authors preregistered showed that the low-overperception network also yielded significantly higher collective outrage ratings than the median perceived outrage of the individual messages. Thus, both networks produced an exaggeration effect.

Study 5 examined whether exposure to a high-overperception newsfeed would shape participants' social evaluations across three domains: perceived social appropriateness of others' posts, perceived ideological temperature (ingroup/outgroup warmth), and perceived ideological extremity of people in their hypothetical social network. Participants who viewed the high-overperception feed rated others' behavior as less socially appropriate and their networks as both warmer toward like-minded others and colder toward political opponents, indicating sharper affective polarization. Finally, they judged their networks as significantly more ideologically extreme than did participants in the low-overperception condition. These results are noteworthy because it suggests that overperceptions of outrage may contribute to perceptions of ideological extremity (Rathje et al., 2021; Van Bavel et al., 2021).

Although the original article did not examine potential partisan differences, it is plausible that Republicans and Democrats would respond differently to overperceived social media feeds. For instance, if Republicans are generally more attuned to ideological threat and perceive higher levels of polarization or hostility in the political environment (Brandt et al., 2014; Crawford, 2017), they might view a high-overperception feed as confirmation of that extremity. Alternatively, Democrats might be more sensitive if the feed appears to show ideological extremity by out-group authors. Because Studies 4 and 5 both tied participants' network exposure to their own political ingroup, it seems theoretically important and design-demanded to model it.

To test this, I ran a linear model predicting collective outrage in Study 4 from condition and participant party. The model was significant, $F(3, 514) = 80.01$, $p < .001$, $R^2_{\text{adj}} = .31$, and the moderating effect of political ideology on condition was significant, $b = -1.30$, $SE = 0.21$, $p < .001$, 95% CI [-1.71, -0.89]. Simple effects showed that Democrats reported similar levels of collective outrage across conditions (High: $M = 1.23$, $SE = 0.10$; Low: $M = 1.31$, $SE = 0.10$), $t(514) = -0.57$, $p = .570$. In contrast, Republicans displayed a substantial difference between conditions, with individuals in the high-overperception condition judged the network as more outraged ($M = 0.45$, $SE = 0.10$) than in the low-overperception condition ($M = -0.77$, $SE = 0.11$), $t(514) = 7.97$, $p < .001$.

Likewise, a linear mixed-effects model was conducted predicting perceived appropriateness of tweets from the tweet's outrage level (Low vs. High), the type of newsfeed (Low vs. High Overperception), and tweet partisanship (Democratic vs. Republican), with random intercepts for participant and tweets for Study 5. There was a significant main effect of outrage level, such that high-outrage tweets were rated as less appropriate than low-outrage tweets, $b = -1.40$, $SE = 0.39$, $t(15.3) = -3.62$, $p = .002$, 95% CI = [-2.12, -0.69]. This effect was moderated by the newsfeed condition, $b = 1.07$, $SE = 0.07$, $t(8213.44) = 14.36$, $p < .001$, 95% CI = [0.93, 1.22], indicating that exposure to a high-overperceived newsfeed increased the perceived appropriateness of outrage

tweets by about one scale point. There was also a significant three-way interaction among newsfeed condition, outrage level, and tweet partisanship, $b = -0.89$, $SE = 0.11$, $t(8213.55) = -7.84$, $p < .001$, 95% $CI = [-1.11, -0.67]$.

A linear model predicting perceived ideological extremity from condition and party affiliation revealed main effects for both participants in the high-overperception condition judged greater extremity, and Republicans also judged greater extremity overall, $b = 0.21$, $SE = 0.07$, $t(960) = 3.01$, $p = .003$. The interaction between condition and political party was not statistically significant, $b = -0.11$, $SE = 0.10$, $t(960) = -1.14$, $p = .260$, indicating that both Democrats and Republicans responded similarly to the newsfeed manipulation in terms of increased perceived extremity. Important for the robustness of the study, the main effect of overperceived feeds remained significant, $b = 0.61$, $SE = 0.07$, $t(960) = 9.09$, $p < .001$, where high overperceived feeds show more perceived extremism than low, $M_{Diff} = .55$, $SE_{Diff} = .05$, $t(960) = 11.28$, $p < .001$.

For Democratic tweets, participants in the low-overperception condition judged high-outrage tweets as less appropriate than low-outrage tweets, $b = -1.40$, $SE = 0.39$, $t(15.30) = -3.62$, $p = .002$. However, this difference was not significant in the high-overperception condition, $b = -0.33$, $SE = 0.39$, $t(15.23) = -0.85$, $p = .407$, indicating that exposure to outrage-saturated newsfeeds normalized expressions of outrage.

In contrast, for Republican tweets (and thus, Republican participants), outrage tweets were judged as less appropriate in both the low, $b = -1.57$, $SE = 0.41$, $t(15.36) = -3.81$, $p = .002$, and high-overperception conditions, $b = -1.38$, $SE = 0.41$, $t(15.34) = -3.36$, $p = .004$. Moreover, exposure to a high-overperception feed increased the perceived appropriateness of Democratic outrage tweets, $b = -1.20$, $SE = 0.09$, $t(1320.28) = -12.67$, $p < .001$. While the higher-overperceived feed also increased the acceptability for Republicans, $b = -0.24$, $SE = 0.11$, $t(1528.26) = -2.26$, $p = .024$, its effect was

much smaller and failed to cancel out the penalties provided for displaying outrage above. Although Republicans consistently judged expressions of outrage as less appropriate and perceived their social networks as more ideologically extreme, they nonetheless responded with greater outgroup coldness when exposed to the outrage-saturated feeds.

4. Conclusion

Our computational reproduction confirms all primary findings reported in Brady et al. (2023). Across five studies, participants consistently overperceived moral outrage, and exposure to outrage-saturated newsfeeds increased perceived collective outrage, reduced social appropriateness, and heightened perceived ideological extremity. All effects replicated in direction, magnitude, and statistical significance, with bootstrapped and/or subsampled refitted confidence intervals nearly identical to the original estimates.

This reproduction reinforces the robustness of the overperception effect and shows that it is insensitive to sampling differences, model specification, or clustering by tweet set. The variable of confidence in one's ratings poses a lot of interesting questions that are beyond the scope of this commentary, as it was not heavily mentioned within the preregistration of the target article. Future individuals interested in exploring this rich dataset could investigate how confidence in one's perception of outrage may increase overperceptions (Carlson et al., 2010) and lead to further polarization, and if this also differs by political ideology.

References

Brady, W. J., McLoughlin, K. L., Torres, M. P., Luo, K. F., Gendron, M., & Crockett, M. J. (2023).

Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility. *Nature Human Behaviour*, 7(6), 917–927. <https://doi.org/10.1038/s41562-023-01582-0>

- Brandt, M. J., Reyna, C., Chambers, J. R., Crawford, J. T., & Wetherell, G. (2014). The ideological-conflict hypothesis: Intolerance among both liberals and conservatives. *Current Directions in Psychological Science*, 23(1), 27–34. <https://doi.org/10.1177/0963721413510932>
- Carlson, E. N., Furr, R. M., & Vazire, S. (2010). Do we know the first impressions we make? Evidence for idiographic meta-accuracy and calibration of first impressions. *Social Psychological and Personality Science*, 1(1), 94–98. <https://doi.org/10.1177/1948550609356028>
- Crawford, J. T. (2017). Are conservatives more sensitive to threat than liberals? It depends on how we define threat and conservatism. *Social Cognition*, 35(4), 354–373. <https://doi.org/10.1521/soco.2017.35.4.354>
- Pek, J., & Wu, H. (2015). Profile likelihood-based confidence intervals and regions for structural equation models. *Psychometrika*, 80(4), 1123–1145. <https://doi.org/10.1007/s11336-015-9461-1>
- Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(26), e2024292118. <https://doi.org/10.1073/pnas.2024292118>
- Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K. C., & Tucker, J. A. (2021). Political psychology in the digital (mis)information age: A model of news belief and sharing. *Social Issues and Policy Review*, 15(1), 84–113. <https://doi.org/10.1111/sipr.12077>
- Willroth, E. C., & Atherton, O. E. (2024). Best laid plans: A guide to reporting preregistration deviations. *Advances in Methods and Practices in Psychological Science*, 7(1), 25152459231213802. <https://doi.org/10.1177/25152459231213802>