Promoting civil discourse on social media using nudges: A tournament of seven interventions

Tatiana Celadin[*,†] Folco Panizza [*,‡] Valerio Capraro [§]

Abstract

In this paper, we test and compare several message-based nudges designed to promote civil discourse and reduce the circulation of hate speech. We conducted a large pre-registered experiment (N = 4,081) to measure the effectiveness of seven nudges: making descriptive norms, injunctive norms, or personal norms salient, cooling down negative emotions, stimulating deliberation or empathy, and highlighting reputation. We used an online platform that reproduces a social media newsfeed and presented the nudge as a message when entering the platform. Our findings indicate that nudges making descriptive norms salient selectively increase participants' overall engagement with relatively harmless content. Additionally, making injunctive norms salient increased the likelihood of liking harmless posts. Exploratory text analysis also reveals that highlighting reputation leads to more substantial and coherent comments on harmful posts. These results suggest that nudges that activate norm considerations represent a promising approach to promoting civil discourse and making social media a safer and more inclusive space for all.

[*] These authors contributed equally to this work.

[†] Department of Economics, University Ca' Foscari of Venice, San Giobbe, Cannaregio 873, 30121, Venice, Italy. Email: tatiana.celadin@unive.it

[‡] IMT School for Advanced Studies Lucca, 55100, Lucca, Italy, Department of Psychology. Email: folco.panizza@imtlucca.it

[§] University of Milan Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126, Milan, Italy. Email: valerio.capraro@unimib.it

1 Introduction

Hate speech has emerged as a pressing issue in contemporary society, especially with the widespread adoption of social media platforms (Avalle et al., 2024; Crockett, 2017). Typically directed at individuals based on characteristics such as ethnicity, sexual orientation, gender, social class, and physical appearance (Silva et al., 2016), hate speech, although perpetrated by only a minority of users (Siegel et al., 2021), can have severe consequences on the well-being of individuals. It can contribute to mental health issues, generating anxiety and fear (Hinduja and Patchin, 2007; Tynes et al., 2008), and leading to social isolation (Siegel, 2020), and can fuel discrimination and prejudice (Müller and Schwarz, 2023), thereby contributing to wider social divisions and conflicts (Henson et al., 2013). As also highlighted by the United Nations, it is of critical importance to identify interventions that can promote civil discourse and create safe spaces where users can express their ideas without fear of discrimination or harm (UN, 2019).

Two widely used approaches to combat online hate speech are content moderation and counter-speech (Siegel, 2020; Windisch et al., 2022). Moderation involves banning, suspending, or hiding comments and profiles that violate the terms and conditions of online platforms. Evidence suggests that content moderation can lead to a reduction of hateful content on those platforms (Tyler et al., 2021; Álvarez-Benjumea and Winter, 2020; Chandrasekharan et al., 2017). However, despite its effectiveness, moderation faces some limitations: banned users may migrate to less regulated platforms, and some hate speech may go undetected by the platforms' algorithms (Parker and Ruths, 2023; Chancellor et al., 2016). On the other hand, counter-speech involves responding to hate speech with positive messages that aim to reduce negative behaviors (Garland et al., 2022). There is evidence that counter-speech does make users less prone to post harmful content (Munger, 2017; Siegel and Badaan, 2020; Hangartner et al., 2021). Yet, this approach is not easily scalable, and the most vulnerable targets of hate may be in a weak position to respond effectively (Tirrell, 2018).

A class of interventions with the potential to overcome some limitations of moderation and counter-speech involves nudging. Nudging offers a less invasive alternative to content moderation, as it does not alter users' material payoffs. Moreover, it is more easily scalable than counter-speech because it relies on architectural changes within the platform (Windisch et al., 2022; Katsaros et al., 2022). Typically, nudging includes adjusting website architecture to promote prosocial behaviors online, such as compliance with the community norms and discouragement of bullying (Grüning et al., 2024; Kraut and Resnick, 2012; Stroud et al., 2017). A particularly relevant form of nudging, closely aligned with our research, involves the use of targeted messages. These message-based nudges are cost-effective, requiring minimal implementation resources while potentially exerting a significant impact. For instance, displaying normative information about community rules during user interactions increased rule compliance in a large-scale field experiment (Matias, 2019). Nudges that elicited empathy and notified users about the potential implications of their posts were effective at promoting empathic responses to instances of cyberbullying (Taylor et al., 2019). Informing users about the audience for their actions enhanced their sense of responsibility and increased the likelihood of flagging cyberbullying posts (DiFranzo et al., 2018).

A crucial step towards the implementation of nudging interventions is the assessment of their relative impact; while specific nudges may be effective, conducting cost-benefit analyses is essential to understanding the most promising interventions (Milkman et al., 2021). Yet, to our knowledge, there is a shortage of studies comparing different nudges aimed at mitigating hate speech in the same social media environment. Our study aims to reduce this gap by examining the efficacy of seven distinct message-based nudges. Following the emerging literature using intervention tournaments (Rendell et al., 2010; Vlasceanu et al., 2024) or megastudies (Milkman et al., 2021; Voelkel et al., 2022; Zickfeld, Ścigała, Elbæk, Michael, Tønnesen, Levy, Ayal, Thielmann, Nockur, Peer, et al., Zickfeld et al.), we test multiple interventions on the same participants pool, allowing a uniform comparison of their relative impacts.

To this end, we selected seven message-based nudges aimed at promoting civil behavior: making descriptive norms, injunctive norms, or personal norms salient, cooling down negative emotions, stimulating deliberation or empathy, and highlighting reputation. We focused on these specific nudges because an extensive body of literature has shown that they can promote prosocial behavior in a broad variety of different social contexts. For example, descriptive norms nudges can increase vaccination rates and promote advocacy for vaccination (Belle and Cantarelli, 2021). Injunctive norms nudges can increase the reporting rates of fake news (Gimpel et al., 2021). Making personal norms salient can heighten generosity and cooperative behavior, with effects persisting in subsequent decisions (Capraro et al., 2019). Encouraging deliberation can increase intentions to wear a face mask during a pandemic (Capraro and Barcelo, 2021). Inducing empathy can favor social distancing and mask wearing (Pfattheicher et al., 2020) as well as positive attitudes toward political outgroups (Masullo, 2023). Social motivations such as reputation increase the accuracy of the assessment of online information (Ronzani et al., 2024; Rathje et al., 2023). Alerts about potentially emotionally charged content can decrease the likelihood of sharing offensive material (Katsaros et al., 2022). In sum, these nudges have been proven to encourage prosocial behavior in various contexts, including in, but not limited to, social media interactions. Therefore, given that reducing engagement with harmful posts and increasing engagement with non-harmful posts can be considered forms of prosocial behavior, we hypothesized that each of these interventions could interact with the harmfulness of the content in determining engagement levels. We do not have a priori hypotheses on their relative effectiveness; the objective of this study is to determine which intervention is most effective.

We recruited 4,081 participants living in the USA through the online recruiting platform Prolific and randomly assigned them to one of eight conditions, including seven message-based nudge interventions and a no-intervention baseline. Participants in all conditions except the baseline were presented with a message, similar to those used in previous studies (Levine et al., 2018). The exact wording of the nudges can be found in Table 1. Then, all participants were redirected to a platform, called Mock Social Media Website Tool, that faithfully reproduces Facebook's newsfeed (Jagayat et al., 2021). This choice aims at increasing the study's ecological validity, in line with previous work (Taylor et al., 2019; DiFranzo et al., 2018; Bhandari et al., 2021; Masur et al., 2021). Participants in each condition interacted with the newsfeed in a manner akin to real social media usage, scrolling through various posts with the option to engage by sharing, commenting, or reacting. Detailed visualization of the newsfeed can be found in Supplementary Material, section H. In the newsfeed, participants were shown 14 posts of varying degrees of harmfulness randomly drawn from a larger pool of 49 posts. The level of harmfulness of each post was determined through an out-of-sample survey where 201 participants rated each post on a scale from 0 to 10 (see Fig. 1 for examples, and osf.io/tsxk2 for the full collection of posts). The mean harmfulness of the posts was 3.68 (SD = 2.72), with a minimum of 0.17 and a maximum of 9.67. The topics for these posts spanned a range of contentious issues, including: abortion, assisted suicide, gun control, marijuana legalization, politics, science (animal testing, climate change, stem cells, vaccination), and social justice (gender equality, LGBTQIA+, racism). Further details about the experimental procedure can be found in the Methods. Descriptive statistics, including means, standard deviations, minima, and maxima for each condition and reaction, are reported in Supplementary Material (SM), Table 1.

Nudge	**Wording**
Descriptive norm	Sometimes people make decisions taking into account what they believe other people would do in the same context. Other times, people make decisions by ignoring what they believe others would do. Many people believe that considering others' expected actions leads to good decision-making. **When we take into account what others would do, we make decisions that are typically socially accepted and widespread.** Please **make your decisions** on this social media platform by **taking into account what you believe others would do**.
Injunctive norm	Sometimes people make decisions taking into account what other people would approve or disapprove of. Other times, people make decisions by ignoring what others consider to be the right thing to do. Many people believe that considering what others approve or disapprove of leads to good decision-making. **When we take into account what others approve or disapprove of, we make decisions that are typically well-regarded.** Please **make your decisions** on this social media platform by **taking into account what you believe others would approve or disapprove of**.
Personal norm	Sometimes people make decisions taking into account what they think is the morally right thing to do. Other times, people make decisions by ignoring their internal sense of right and wrong. Many people believe that considering their internal morality leads to good decision-making. **When we take into account what we believe to be the right thing, we make decisions that are typically in line with our deepest beliefs**. Please **make your decisions** on this social media platform by **relying on what you think is the morally right thing to do**.
Negative emotions	Sometimes people make decisions following their immediate negative emotions. Other times, people make decisions by letting their emotions cool down first. Many people believe that avoiding their immediate negative emotions leads to good decision-making. **When we avoid our immediate negative emotions, we make decisions that typically prevent us from feeling bad.** Please **make your decisions** on this social media platform by **letting your negative emotions cool down**.
Deliberation	Sometimes people make decisions taking into account what they think is the rational thing to do. Other times, people make decisions by ignoring their logic and reason. Many people believe that considering their rational side leads to good decision-making. **When we take into account our analytic part, we make decisions that are typically well-thought.** Please **make your decisions** on this social media platform by **taking into account what you think is the rational thing to do**.
Empathy	Sometimes people make decisions taking into account the point of view of the other people involved. Other times, people make decisions by ignoring the point of view of others. Many people believe that putting oneself in the shoes of others leads to good decision-making. **When we take into account what other people experience from their perspective, we make decisions that are typically empathic.** Please **make your decisions** on this social media platform by **taking into account the point of view of others**.
Reputation	Sometimes people make decisions taking into account how these decisions will affect their own reputation. Other times, people make decisions by ignoring their effect on reputation. Many people believe that considering how their decisions will impact their own reputation leads to good decision-making. **When we take into account that our actions are judged by others, we make decisions that are typically well-evaluated.** Please **make your decisions** on this social media platform by **taking into account your reputation**.

Table 1: Exact wording of the seven message-based nudges. Messages were displayed on the screen just before participants entered the social media newsfeed.

2 Results

We begin by examining the overall engagement, defined as the sum of all possible actions (reactions, comments, shares) taken by the participants. As pre-registered, we conduct a linear regression with robust standard errors clustered at the participant and post levels. As regressors, we include the harmfulness of the post, seven dummies, one for each intervention, and the seven interactions between each of the intervention dummies and the harmfulness of the post.

Our key variables of interest are the seven interactions, which measure how the difference between overall engagement in the corresponding intervention and overall engagement in the baseline varies when the harmfulness of the post increases. We find that this interaction is significant and negative when the nudge is based on descriptive norms ($\beta = -0.011$, $t = -3.66$, $p = 0.001$), injunctive norms ($\beta = -0.007$, $t = -2.63$, $p = 0.012$), deliberation ($\beta = -0.006$, $t = -2.07$, $p = 0.044$), and cooling down negative emotions ($\beta = -0.009$, $t = -3.23$, $p = 0.002$). The effects of descriptive norms and cooling down negative emotions are robust to Bonferroni correction.

These results show that as the harmfulness of posts increases, the differences between overall engagement in the interventions and overall engagement in the baseline decrease. This trend could be explained by one of two non-mutually exclusive mechanisms: (i) the interventions decrease engagement with relatively harmful posts, or (ii) the interventions increase engagement with relatively harmless posts. To determine which of these two mechanisms is at work, as an exploratory analysis, we look at the simple effects of the interventions. This analysis allows us to estimate the interventions' effects separately for posts with extreme values of harmfulness.

For extremely harmless posts, the model estimates significantly higher engagement in all interventions, compared to the baseline, with the exception of the reputation nudge. This increase is robust to Bonferroni correction for personal ($\beta = 0.059$, $t = 3.09$, $p = 0.003$), descriptive ($\beta = 0.114$, $t = 4.99$, $p < 0.001$) and injunctive ($\beta = 0.078$, $t = 3.88$, $p < 0.001$) norm nudges. According to the model's predictions, for every ten posts, participants in the baseline condition engage an average of 1.65 times (95% CI = $[1.28, 2.01]$). The engagement rate increases to 2.24 times (95% CI = $[1.84, 2.63]$) with a personal norm message, 2.42 times (95% CI = $[2.02, 2.83]$) with an injunctive norm message, and 2.77 times (95% CI = $[2.26, 3.29]$) with a descriptive norm message.

For extremely harmful posts, engagement levels appear to be closely aligned across conditions. For instance, the model estimates 1.15 engaged posts per ten in the baseline scenario (95% CI = $[0.72, 1.58]$), closely matched by 1.14 engaged posts in the descriptive norm condition (95% CI = $[0.66, 1.62]$).

These analyses suggest that the channel through which the descriptive norm intervention works is by increasing overall engagement with relatively harmless posts rather than decreasing engagement with more harmful posts. On the other hand, the cooling down negative emotions intervention likely works through a combination of two non-significant effects: an increase in engagement with harmless posts and a decrease in engagement with harmful posts. These conclusions are exemplified in Fig. 1, 'engagement score' panel, where, for simplicity, we categorized the posts into three groups, according to their level of harmfulness.

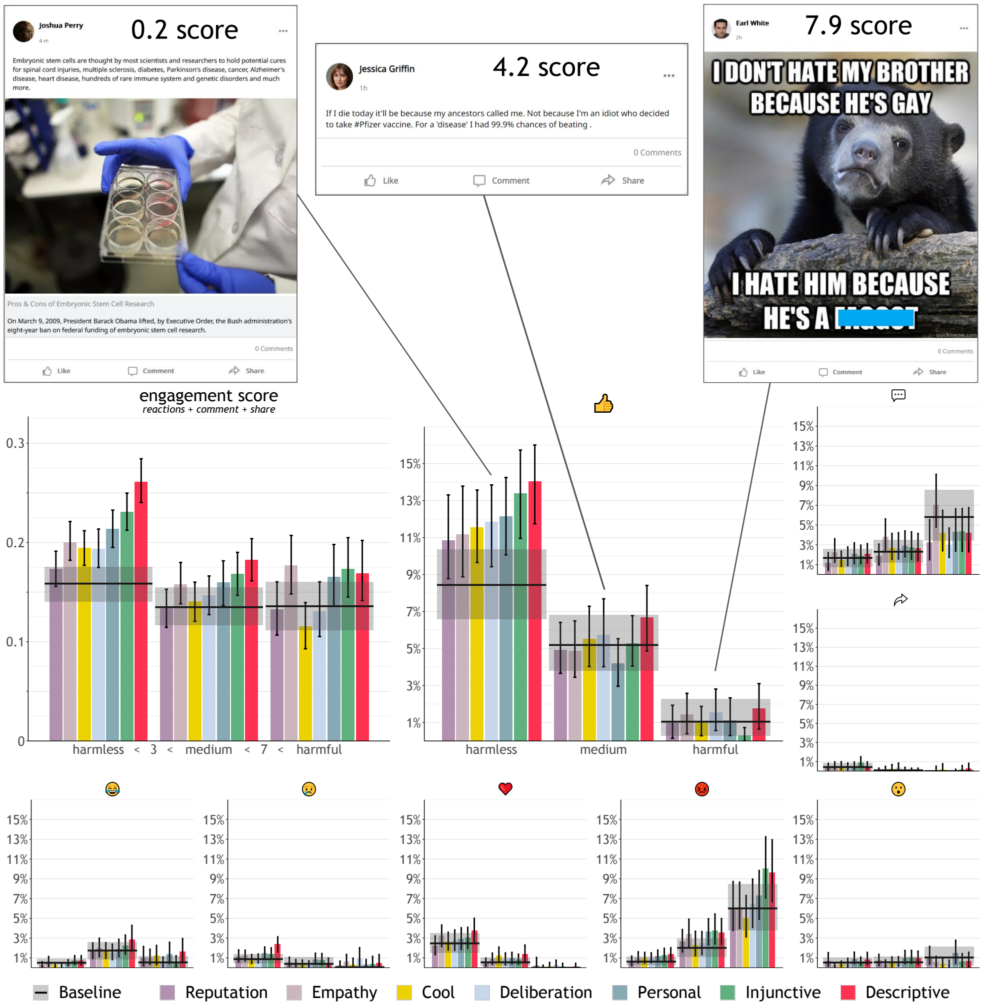
Next, as pre-registered, we examine the individual reactions to understand which reactions drive these changes in engagement, starting with the most common reaction: liking a post. To this end, we conduct a logit regression with robust standard errors clustered at the participant and post levels. As before, we include the harmfulness of the post, seven dummy variables for each intervention, and seven interactions between each intervention dummy and the harmfulness of the post as regressors. Our key variables of interest are the seven interactions. We find significant negative interactions for nudges based on personal norms ($\beta = -0.122$, $z = -2.49$, $p = 0.013$), descriptive norms ($\beta = -0.073$, $z = -2.07$, $p = 0.039$), and injunctive norms ($\beta = -0.160$, $z = -4.02$, $p < 0.001$). The effect of nudging the injunctive norms is robust to Bonferroni correction.

Examining the simple effects, participants in the baseline condition are predicted to like an average of 1.27 out of every ten harmless posts (95% CI = $[1.12, 1.43]$), compared to 2.16 posts in the injunctive norm intervention (95% CI = $[1.96, 2.36]$). For extremely harmful posts, the predicted liking rates are very low across all conditions and never significant after Bonferroni correction. This suggests that the injunctive norm intervention works primarily by increasing the liking of harmless posts rather than decreasing the liking of harmful ones. See Fig. 1, 'like' panel.

We then investigate the other reactions – love, laugh, anger, cry, and wow – as well as commenting and sharing. Using logit regressions, we account for the harmfulness of the post, intervention groups, and their interactions. We observe a general decrease in 'love' reactions, juxtaposed with a rise in reactions of 'anger', 'wow', and in commenting behavior (center-right and bottom center in Figure 1). This shift from positive to more contentious or surprised reactions serves as a compensatory mechanism, preventing engagement levels from plummeting. On the other hand, the interactions between the intervention dummies and harmfulness are never significant after Bonferroni correction. In sum, nudging interventions affect engagement primarily through changes in the frequency of 'likes'. We refer to the SM, section 2, for full regression tables.

As a robustness check, we investigate whether variables such as the topic of the post, gender, age, and political orientation might moderate the efficacy of the interventions. We found significant moderation only in the context of the post's topic. Specifically, the effectiveness of the intervention aimed at cooling down negative emotions was predominantly evident in posts concerning assisted suicide and gun control. Similarly, the descriptive norm nudge showed a more pronounced influence on posts related to assisted suicide, gun control, legalization, and politics. See SM, sections 3-6, for regression tables.

As the final phase of the analysis, we conducted a series of exploratory text analyses to test for differences in comment style across conditions. To analyze this textual data, we use PeRspective API, a tool developed by Google Jigsaw, which uses pre-trained machine learning algorithms to analyze conversational content (Lees et al., 2022). Each comment (total $N = 1,507$) was thus weighted against a variety of metrics, and normalized on a scale from 0 to 1 (see SM, section 7, for the entire list of metrics including a definition for each). We run a series of mixed-effects linear regressions, one for each metric, treating these metrics as dependent variables. The models include the seven intervention dummies, the level of harmfulness of the post, and their interaction, as predictor variables. Standard errors are clustered at the participant and post levels. Once again, our variables of interest are the interaction terms. For two specific metrics, substance, and coherence, we found these interactions to be significant in most interventions (see SM, Tables 7-9, for regression tables). For both metrics, the interaction is robust to Bonferroni correction only in the reputation condition (substance: $\beta = 0.249$, $t = 3.86$, $p < .001$; coherence: $\beta = 0.298$, $t = 3.70$, $p < .001$). Note, however, that these two metrics are strongly correlated (Pearson correlation coefficient: $r(1563) = 0.312$, $t = 12.95$, $p < 0.001$). Essentially, as the harmfulness of a comment increases, participants subjected to the reputation condition are more likely to leave more substantial and coherent responses compared to participants in the baseline condition. An illustrative example of a substantial and coherent comment left by one participant is: ‘I’m not necessarily an advocate of any mind altering substances, however when you break it down to risks, health concerns, and other aspects I feel like the laws to allow people to drink alcohol but not consume marijuana contradict themselves.’).



Joshua Perry

4 m

0.2 score

Embryonic stem cells are thought by most scientists and researchers to hold potential cures for spinal cord injuries, multiple sclerosis, diabetes, Parkinson's disease, cancer, Alzheimer's disease, heart disease, hundreds of rare immune system and genetic disorders and much more.

Pros & Cons of Embryonic Stem Cell Research

On March 9, 2009, President Barack Obama lifted, by Executive Order, the Bush administration's eight-year ban on federal funding of embryonic stem cell research.

0 Comments

Like Comment Share

Jessica Griffin

1h

4.2 score

If I die today it'll be because my ancestors called me. Not because I'm an idiot who decided to take #Pfizer vaccine. For a 'disease' I had 99.9% chances of beating .

0 Comments

Like Comment Share

Earl White

2h

7.9 score

I DON'T HATE MY BROTHER BECAUSE HE'S GAY

I HATE HIM BECAUSE HE'S A [redacted]

0 Comments

Like Comment Share

Figure 1: Top: three sample posts from the newsfeed, each annotated with its harmfulness score. The content of the highly harmful post is obscured but was fully readable during the experiment. Bottom, clockwise from the top left: composite engagement score, frequency of 'like' reactions, and frequency of the other reactions. These metrics are averaged across both participants and posts with comparable ranges of harmfulness. Posts are divided into three levels for illustrative purposes. In the regressions, harmfulness is treated as a continuous variable. The colored bars represent the average metrics for each experimental condition, whereas the black horizontal lines denote the baseline averages. Error bars and the shaded regions around the averages represent confidence intervals adjusted for multiple comparisons using Bonferroni-corrected bootstrap methods.

3 Discussion

We tested seven message-based nudges designed to combat the spread of hate speech on social media. Our findings indicate that a nudge making descriptive norms salient increases participants' overall engagement with harmless content and that a nudge making injunctive norms salient increases the likelihood of participants liking harmless posts.

Social media interventions may work through two distinct, although not mutually exclusive, mechanisms: reducing interactions with harmful content or boosting interactions with harmless content. Since the vast majority of online content is harmless, some scholars have argued that increasing engagement with harmless content is as important, if not more so, than reducing engagement with harmful content. This is because the ratio of harmless to harmful content, which is the essential factor defining the overall quality of online content, would be more strongly impacted (Álvarez-Benjumea and Winter, 2020; Capraro and Celadin, 2023). From this perspective, one of the positive aspects of these interventions is that they work precisely through this mechanism.

Understanding why these interventions appear to work primarily through this mechanism is an interesting direction for future work. At this stage of research, we can only speculate. Looking at Figure 1, one may notice some promising trends for harmful content. The personal, descriptive, and injunctive norm nudges seem to increase the angry reaction to harmful posts. Moreover, most interventions appear to reduce the frequency of comments. However, these trends do not reach common thresholds of statistical significance. This may be due to the limited power of this study to detect significant effects for less used reactions. In other words, it is possible that the null effects of the interventions on harmful posts stem from the combination of two 'socially positive' effects: one that leads people to react more angrily to harmful posts, and another that encourages people to ignore harmful posts and avoid commenting on them. Future experiments with a much larger sample size can illuminate this point. Regardless of the outcomes of these experiments, it is important to note that the overall positive effect on engagement of the descriptive norm and cooling down negative emotions interventions is promising from a practical perspective. It has been argued that social media platforms have a tendency to maximize engagement, even at the cost of promoting harmful content (Stacey and Bradshaw, 2021). From this perspective, it is encouraging that these interventions increase engagement while promoting harmless content.

We also found that the nudge aimed at cooling down negative emotions interacts with the harmfulness of the posts in the predicted direction. The analysis of simple effects provides evidence that this intervention likely operates through a combination of two mechanisms: increasing engagement with harmless posts and decreasing engagement with harmful posts. However, none of these effects was singularly significant, possibly due to the limited power of our study. Future work could investigate more thoroughly the capacity of this specific nudge intervention to symmetrically affect engagement for both harmful and harmless posts.

A strength of message-based interventions lies in their scalability, which stands in contrast to the resource-intensive nature of counter-speech strategies and content moderation by human reviewers. Message-based nudges can be easily integrated via architectural changes within a platform. Furthermore, their implementation can be recurrent, using reminders like pop-ups when users return on a social media platform after a period of inactivity. Nonetheless, message-based nudges are not without their limitations. One concern regards their modest impact, especially when compared to more significant structural modifications to a platform. For instance, one study demonstrated that introducing a button to flag misinformation reduced the sharing of such content by 25%, whereas an accuracy nudge resulted in only a 5% decrease (Pretus et al., 2024). In this regard, it is important to note that the effect sizes for the most successful interventions in our study were substantial. The total engagement rate in the descriptive norm condition was 68% higher than in the baseline. Similarly, the average number of likes in the injunctive norm condition rose by 70%, compared to the baseline.

Another set of concerns involves the possibility that the effectiveness of message-based nudges may decrease over time (Zeng et al., 2023). Moreover, especially if the nudges are repeated too frequently, there is the potential for user desensitization. Future work should explore the boundary conditions of these specific nudges, bearing in mind that addressing a complex challenge like the reduction of hate speech likely requires more than a single type of intervention. Message-based nudges should not be regarded as the definitive solution, but as one tool among many in a comprehensive strategy aimed at mitigating hate speech. From this perspective, another promising avenue for future research is exploring how message-based nudges can be employed synergistically with other interventions to create a more cohesive and effective approach.

Our conclusive exploratory text-analysis revealed that participants exposed to a message about the consequences for their reputation tended to write more substantial and coherent comments, in response to harmful posts. This trend was observed also in several other interventions, but it was not robust to Bonferroni correction. It is important to acknowledge, however, that these findings, derived from a subset of participants, may not be sufficiently powered to draw definitive conclusions. Moreover, the metrics for evaluating the substance and coherence of comments were originally trained on the content of a single newspaper (The New York Times) and may be biased by the readership of that journal. Additionally, due to technical constraints, some lengthy messages ($N = 36$) exceeding 255 characters were truncated in our dataset. In this case, the analyses were based solely on the available portions of these messages. Future research should employ preregistered, more powerful designs to investigate in greater detail how message-based nudges influence commenting style.

Overall, these results suggest that some message-based nudges, and in particular those activating normative considerations, could help create a more positive and inclusive online environment. Future work should investigate the mechanisms through which these interventions work and their boundary conditions.

4 Methods

This study was approved by the Middlesex University Ethics Committee n. 21556. Pre-registration and data are available at: osf.io/tsxk2

We selected 71 posts from different platforms (e.g., Facebook, Twitter, Reddit, 4chan). We recruited 201 participants through Prolific to rate the harmfulness of these posts along two dimensions:

- "How abusive do you think this post is?";
- "How hateful do you think this post is?".

Since the two dimensions were consistently correlated (mean by-post Cronbach's $\alpha = 0.80$), we aggregated them into a single *harmfulness* index. Some posts were excluded to make the levels of harmfulness as heterogeneous as possible. In addition, some posts were discarded because there were too many conservative-leaning posts with high values of harmfulness. The exclusion was done by random sampling so that the selection of posts could not be biased by the researchers. Due to an error in pairing posts with users, a post describing an abortion experience was mistakenly paired with a male avatar. The error was discovered after data collection had begun and resulted in the exclusion of 606 participants who viewed the post, leading to the second data collection. Thus, the final set of stimuli contained 49 posts: 27 were conservative, 22 were progressive; on a scale of 0 to 10, the mean harmfulness was 3.68 (SD = 2.72), with a minimum of 0.17 and a maximum of 9.67.

Once we collected the harmfulness ratings of the posts, we ran the main experiment. We recruited 4,081 participants from the USA through Prolific and randomly assigned them to one of eight conditions, including seven nudges and the baseline. Specifically, we ran two sessions. In the first session (Sept 12, 2022), we collected N=1,442 subjects, and in the second session (Oct 6-8, 2022), we collected N=2,639 subjects. Participants were shown 14 posts of varying degrees of harmfulness. To increase ecological validity, we used a new platform, called Mock Social Media Website Tool, that faithfully reproduces Facebook's newsfeed (Jagayat et al., 2021). Names and profile pictures provided in the Facebook newsfeed were randomly generated online and for research purposes only through behindthename.com/random and generated.photos. The macro-topics of the posts were: abortion, assisted suicide, gun control, legalisation, politics, science, and social justice. The science macro-topic included posts on animal testing, climate change, stem cell, and vaccination; the social justice macro-topic included posts on gender equality, LGBTQIA+, and racism. Participants in each condition could interact with the posts by sharing, commenting, or reacting to them. Before accessing the newsfeed, all the conditions except the baseline presented participants with a nudge message. The messages can be found in Table 1.

Acknowledgments and Funding

This research was funded by the RKEFF Middlesex University Research Grant.

Author contributions statement

The authors declare no competing interest. TC, FP, VC developed the research; FP designed the survey; TC, FP, VC analyzed data; TC, VC wrote the paper; TC, FP, VC revised the paper.

Data availability

Pre-registration, data, analysis code, and materials have been deposited in https://osf.io/tsxk2/. ChatGPT was used to polish the text.

References

Álvarez-Benjumea, A. and F. Winter (2020). The breakdown of antiracist norms: A natural experiment on hate speech after terrorist attacks. *Proceedings of the National Academy of Sciences 117*(37), 22800–22804.

Avalle, M., N. Di Marco, G. Etta, E. Sangiorgio, S. Alipour, A. Bonetti, L. Alvisi, A. Scala, A. Baronchelli, M. Cinelli, et al. (2024). Persistent interaction patterns across social media platforms and over time. *Nature*, 1–8.

Belle, N. and P. Cantarelli (2021). Nudging public employees through descriptive social norms in healthcare organizations. *Public Administration Review 81*(4), 589–598.

Bhandari, A., M. Ozanne, N. N. Bazarova, and D. DiFranzo (2021). Do you care who flagged this post? Effects of moderator visibility on bystander behavior. *Journal of Computer-Mediated Communication 26*(5), 284–300.

Capraro, V. and H. Barcelo (2021). Telling people to "rely on their reasoning" increases intentions to wear a face covering to slow down covid-19 transmission. *Applied Cognitive Psychology 35*(3), 693–699.

Capraro, V. and T. Celadin (2023). "I think this news is accurate": Endorsing accuracy decreases the sharing of fake news and increases the sharing of real news. *Personality and Social Psychology Bulletin 49*, 1635–1645.

Capraro, V., G. Jagfeld, R. Klein, M. Mul, and I. van de Pol (2019). Increasing altruistic and cooperative behaviour with simple moral nudges. *Scientific Reports 9*(1), 11880.

Chancellor, S., J. A. Pater, T. Clear, E. Gilbert, and M. De Choudhury (2016). # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pp. 1201–1213.

Chandrasekharan, E., U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert (2017). You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction 1*(CSCW), 1–22.

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

DiFranzo, D., S. H. Taylor, F. Kazerooni, O. D. Wherry, and N. N. Bazarova (2018). Upstanding by design: Bystander intervention in cyberbullying. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–12.

Garland, J., K. Ghazi-Zahedi, J.-G. Young, L. Hébert-Dufresne, and M. Galesic (2022). Impact and dynamics of hate and counter speech online. *EPJ data science 11*(1), 3.

Gimpel, H., S. Heger, C. Olenberger, and L. Utz (2021). The effectiveness of social norms in fighting fake news on social media. *Journal of Management Information Systems 38*(1), 196–221.

Grüning, D. J., J. Kamin, F. Panizza, M. Katsaros, and P. Lorenz-Spreen (2024). A framework for promoting online prosocial behavior via digital interventions. *Communications Psychology 2*(1), 6.

Hangartner, D., G. Gennaro, S. Alasiri, N. Bahrich, A. Bornhoft, J. Boucher, B. B. Demirci, L. Derksen, A. Hall, M. Jochum, et al. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences 118*(50), e2116310118.

Henson, B., B. W. Reyns, and B. S. Fisher (2013). Fear of crime online? Examining the effect of risk, previous victimization, and exposure on fear of online interpersonal victimization. *Journal of Contemporary Criminal Justice 29*(4), 475–497.

Hinduja, S. and J. W. Patchin (2007). Offline consequences of online victimization: School violence and delinquency. *Journal of School Violence 6*(3), 89–112.

Jagayat, A., B. Gurkaran, P. Carson, and C. L. Becky (2021). Mock social media website tool (1.0). *Computer software https://docs.studysocial.media*.

Katsaros, M., K. Yang, and L. Fratamico (2022). Reconsidering tweets: Intervening during tweet creation decreases offensive content. In *Proc. Int. AAAI Conf. Web Soc. Med.*, Volume 16, pp. 477–487.

Kraut, R. E. and P. Resnick (2012). *Building successful online communities: Evidence-based social design.* MIT Press.

Lees, A., V. Q. Tran, Y. Tay, J. Sorensen, J. Gupta, D. Metzler, and L. Vasserman (2022). A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3197–3207.

Levine, E. E., A. Barasch, D. Rand, J. Z. Berman, and D. A. Small (2018). Signaling emotion and reason in cooperation. *Journal of Experimental Psychology: General 147*(5), 702.

Masullo, G. M. (2023). A new solution to political divisiveness: Priming a sense of common humanity through facebook meme-like posts. *New Media & Society*, 14614448231184633.

Masur, P. K., D. DiFranzo, and N. N. Bazarova (2021). Behavioral contagion on social media: Effects of social norms, design interventions, and critical media literacy on self-disclosure. *Plos one 16*(7), e0254670.

Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences 116*(20), 9785–9789.

Milkman, K. L., D. Gromet, H. Ho, J. S. Kay, T. W. Lee, P. Pandiloski, Y. Park, A. Rai, M. Bazerman, J. Beshears, et al. (2021). Megastudies improve the impact of applied behavioural science. *Nature 600*(7889), 478–483.

Müller, K. and C. Schwarz (2023). From hashtag to hate crime: Twitter and antiminority sentiment. *American Economic Journal: Applied Economics 15*(3), 270–312.

Munger, K. (2017). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior 39*, 629–649.

Parker, S. and D. Ruths (2023). Is hate speech detection the solution the world wants? *Proceedings of the National Academy of Sciences 120*(10), e2209384120.

Pfattheicher, S., L. Nockur, R. Böhm, C. Sassenrath, and M. B. Petersen (2020). The emotional path to action: Empathy promotes physical distancing and wearing of face masks during the covid-19 pandemic. *Psychological Science 31*(11), 1363–1373.

Pretus, C., A. M. Javeed, D. Hughes, K. Hackenburg, M. Tsakiris, O. Vilarroya, and J. J. Van Bavel (2024). The misleading count: an identity-based intervention to counter partisan misinformation sharing. *Philosophical Transactions of the Royal Society B 379*(1897), 20230040.

Rathje, S., J. Roozenbeek, J. J. Van Bavel, and S. van der Linden (2023). Accuracy and social motivations shape judgements of (mis) information. *Nature Human Behaviour*, 1–12.

Rendell, L., R. Boyd, D. Cownden, M. Enquist, K. Eriksson, M. W. Feldman, L. Fogarty, S. Ghirlanda, T. Lillicrap, and K. N. Laland (2010). Why copy others? Insights from the social learning strategies tournament. *Science 328*(5975), 208–213.

Ronzani, P., F. Panizza, T. Morisseau, S. Mattavelli, and C. Martini (2024). How different incentives reduce scientific misinformation online. *Harvard Kennedy School Misinformation Review*.

Siegel, A. A. (2020). Online hate speech. *Social media and democracy: The state of the field, prospects for reform*, 56–88.

Siegel, A. A. and V. Badaan (2020). # no2sectarianism: Experimental approaches to reducing sectarian hate speech online. *American Political Science Review 114*(3), 837–855.

Siegel, A. A., E. Nikitin, P. Barberá, J. Sterling, B. Pullen, R. Bonneau, J. Nagler, J. A. Tucker, et al. (2021). Trumping hate on Twitter? Online hate speech in the 2016 us election campaign and its aftermath. *Quarterly Journal of Political Science 16*(1), 71–104.

Silva, L., M. Mondal, D. Correa, F. Benevenuto, and I. Weber (2016). Analyzing the targets of hate in online social media. In *Proc. Int. AAAI Conf. Web Soc. Med.*, Volume 10, pp. 687–690.

Stacey, K. and T. Bradshaw (2021). Facebook chose to maximise engagement at users' expense, whistleblower says. *Financial Times*.

Stroud, N. J., A. Muddiman, and J. M. Scacco (2017). Like, recommend, or respect? Altering political behavior in news comment sections. *New Media & Society 19*(11), 1727–1743.

Taylor, S. H., D. DiFranzo, Y. H. Choi, S. Sannon, and N. N. Bazarova (2019). Accountability and empathy by design: Encouraging bystander intervention to cyberbullying on social media. *Proceedings of the ACM on Human-Computer Interaction 3*(CSCW), 1–26.

Tirrell, L. (2018). Toxic misogyny and the limits of counterspeech. *Fordham Law Review 87*, 2433.

Tyler, T., M. Katsaros, T. Meares, and S. Venkatesh (2021). Social media governance: Can social media companies motivate voluntary rule following behavior among their users? *Journal of Experimental Criminology 17*, 109–127.

Tynes, B. M., M. T. Giang, D. R. Williams, and G. N. Thompson (2008). Online racial discrimination and psychological adjustment among adolescents. *Journal of Adolescent Health 43*(6), 565–569.

UN (2019). United nations strategy and plan of action on hate speech. *Available at https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml*.

Vlasceanu, M., K. C. Doell, J. B. Bak-Coleman, B. Todorova, M. M. Berkebile-Weinberg, S. J. Grayson, Y. Patel, D. Goldwert, Y. Pei, A. Chakroff, et al. (2024). Addressing climate change with behavioral science: A global intervention tournament in 63 countries. *Science Advances 10*(6), eadj5778.

Voelkel, J. G., M. Stagnaro, J. Chu, S. Pink, J. Mernyk, C. Redekopp, M. Cashman, Q. Submitters, J. Druckman, D. Rand, et al. (2022). Megastudy identifying successful interventions to strengthen americans' democratic attitudes. *Available at https://www.ipr.northwestern.edu/documents/working-papers/2022/wp-22-38.pdf*.

Windisch, S., S. Wiedlitzka, A. Olaghere, and E. Jenaway (2022). Online interventions for reducing hate speech and cyberhate: A systematic review. *Campbell Systematic Reviews 18*(2), e1243.

Zeng, Z., H. Dai, D. J. Zhang, H. Zhang, R. Zhang, Z. Xu, and Z.-J. M. Shen (2023). The impact of social nudges on user-generated content for social network platforms. *Management Science 69*(9), 5189–5208.

Zickfeld, J. H., K. A. Ścigała, C. T. Elbæk, J. Michael, M. H. Tønnesen, G. Levy, S. Ayal, I. Thielmann, L. Nockur, E. Peer, et al. I solemnly swear I'm up to good: A megastudy investigating the effectiveness of honesty oaths on curbing dishonesty. *Available at: https://osf.io/preprints/psyarxiv/hctxe*.

Supplementary Material

A Descriptives

Condition	Angry	Comment	Haha	Like	Love	Sad	Share	Wow	Engagement
Baseline	0.29	0.37	0.14	0.85	0.20	0.07	0.03	0.09	1.94
	(0.88)	(1.34)	(0.57)	(1.61)	(0.65)	(0.35)	(0.25)	(0.49)	(3.30)
	[0,5]	[0,14]	[0,5]	[0,14]	[0,5]	[0,5]	[0,4]	[0,6]	[0,14]
Cool	0.29	0.32	0.13	1.07	0.25	0.09	0.05	0.06	2.13
	(0.92)	(1.14)	(0.50)	(1.87)	(0.75)	(0.40)	(0.35)	(0.34)	(3.47)
	[0,9]	[0,10]	[0,5]	[0,12]	[0,5]	[0,4]	[0,5]	[0,5]	[0,14]
Deliberation	0.36	0.33	0.11	1.08	0.23	0.10	0.05	0.06	2.18
	(1.10)	(1.21)	(0.43)	(1.79)	(0.83)	(0.36)	(0.32)	(0.37)	(3.44)
	[0,9]	[0,13]	[0,3]	[0,14]	[0,8]	[0,4]	[0,3]	[0,6]	[0,14]
Descriptive	0.50	0.36	0.21	1.28	0.32	0.20	0.05	0.12	2.86
	(1.25)	(1.20)	(0.62)	(1.73)	(0.83)	(0.55)	(0.26)	(0.52)	(3.95)
	[0,8]	[0,12]	[0,5]	[0,10]	[0,6]	[0,4]	[0,3]	[0,6]	[0,14]
Empathy	0.40	0.50	0.16	1.02	0.25	0.11	0.04	0.05	2.39
	(1.25)	(1.65)	(0.56)	(1.61)	(0.71)	(0.41)	(0.29)	(0.26)	(3.67)
	[0,10]	[0,14]	[0,5]	[0,13]	[0,5]	[0,4]	[0,3]	[0,3]	[0,14]
Injunctive	0.52	0.36	0.17	1.17	0.24	0.14	0.07	0.11	2.61
	(1.45)	(1.32)	(0.54)	(1.74)	(0.69)	(0.50)	(0.30)	(0.46)	(3.90)
	[0,10]	[0,11]	[0,4]	[0,8]	[0,6]	[0,4]	[0,3]	[0,4]	[0,14]
Personal	0.45	0.39	0.16	1.04	0.24	0.15	0.05	0.14	2.47
	(1.23)	(1.30)	(0.56)	(1.57)	(0.69)	(0.49)	(0.28)	(0.60)	(3.69)
	[0,9]	[0,12]	[0,4]	[0,9]	[0,6]	[0,4]	[0,3]	[0,6]	[0,14]
Reputation	0.34	0.26	0.13	0.98	0.17	0.11	0.04	0.11	2.01
	(1.09)	(1.22)	(0.48)	(1.67)	(0.61)	(0.39)	(0.22)	(0.52)	(3.39)
	[0,11]	[0,13]	[0,4]	[0,14]	[0,6]	[0,3]	[0,2]	[0,6]	[0,14]

Table 2: Descriptive statistics: For each condition and each reaction, we report the mean, the standard deviation (in round brackets), and the minimum and maximum values (in square brackets).

B Main Analysis

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 9
	Engagment	Like	Angry	Laugh	Love	Cry	Wow	Sharing	Comment	Reactions
Harmful	-0.005	-0.272***	0.324***	0.070	-0.521***	-0.150*	0.132*	-0.490***	0.206***	-0.100***
	(0.003)	(0.038)	(0.037)	(0.052)	(0.102)	(0.075)	(0.066)	(0.135)	(0.037)	(0.024)
Personal	0.059**	0.442**	0.667*	-0.127	0.040	0.538*	0.788	0.159	0.289	0.359**
	(0.019)	(0.144)	(0.323)	(0.295)	(0.286)	(0.225)	(0.440)	(0.542)	(0.356)	(0.122)
Cool	0.047*	0.368**	0.450	-0.416	-0.023	-0.162	0.064	0.036	0.176	0.315*
	(0.019)	(0.128)	(0.341)	(0.249)	(0.278)	(0.422)	(0.481)	(0.484)	(0.355)	(0.128)
Reputation	0.017	0.274*	-0.119	-0.410	-0.244	0.292	0.293	0.245	-0.161	0.109
	(0.019)	(0.132)	(0.331)	(0.376)	(0.274)	(0.221)	(0.529)	(0.417)	(0.416)	(0.130)
Descriptive	0.114***	0.596***	0.689*	0.345	0.358	1.116***	0.932	-0.159	0.412	0.654***
	(0.023)	(0.133)	(0.322)	(0.294)	(0.270)	(0.254)	(0.540)	(0.520)	(0.364)	(0.123)
Injunctive	0.078***	0.639***	0.508	0.268	0.093	0.494	0.573	0.481	0.193	0.462***
	(0.020)	(0.137)	(0.281)	(0.225)	(0.257)	(0.348)	(0.455)	(0.432)	(0.331)	(0.125)
Empathy	0.041*	0.311*	0.467	-0.054	0.211	0.240	-0.774	-0.108	0.395	0.257
	(0.020)	(0.157)	(0.296)	(0.398)	(0.267)	(0.218)	(0.470)	(0.444)	(0.357)	(0.132)
Deliberation	0.040*	0.332*	0.298	-0.513	-0.035	-0.289	-0.220	-0.031	0.408	0.218
	(0.020)	(0.136)	(0.303)	(0.301)	(0.339)	(0.380)	(0.583)	(0.530)	(0.370)	(0.124)
Personal×Harmful	-0.005	-0.122*	-0.028	0.064	0.083	0.057	-0.077	0.135	-0.043	-0.014
	(0.003)	(0.049)	(0.039)	(0.042)	(0.107)	(0.069)	(0.077)	(0.176)	(0.048)	(0.024)
Cool×Harmful	-0.009**	-0.065	-0.076	0.072	0.146	0.114	-0.095	0.249	-0.061	-0.054*
	(0.003)	(0.037)	(0.052)	(0.042)	(0.092)	(0.100)	(0.088)	(0.173)	(0.044)	(0.026)
Reputation×Harmful	-0.003	-0.064	0.047	0.070	0.055	0.022	-0.021	-0.079	-0.037	0.003
	(0.003)	(0.038)	(0.040)	(0.057)	(0.103)	(0.062)	(0.076)	(0.179)	(0.055)	(0.024)
Descriptive×Harmful	-0.011***	-0.073*	-0.020	0.017	0.079	-0.044	-0.146	0.320	-0.091*	-0.032
	(0.003)	(0.035)	(0.040)	(0.045)	(0.101)	(0.073)	(0.097)	(0.187)	(0.044)	(0.023)
Injunctive×Harmful	-0.007*	-0.160***	0.025	-0.016	0.041	0.048	-0.074	0.163	-0.037	-0.016
	(0.003)	(0.040)	(0.032)	(0.035)	(0.092)	(0.086)	(0.078)	(0.164)	(0.041)	(0.023)
Empathy×Harmful	-0.002	-0.060	-0.019	0.042	0.002	0.067	0.049	0.247	-0.011	-0.014
	(0.003)	(0.045)	(0.037)	(0.060)	(0.091)	(0.072)	(0.073)	(0.171)	(0.044)	(0.026)
Deliberation×Harmful	-0.006*	-0.037	-0.009	0.055	0.100	0.175	-0.042	0.243	-0.106*	-0.011
	(0.003)	(0.042)	(0.034)	(0.044)	(0.132)	(0.097)	(0.095)	(0.153)	(0.053)	(0.024)
Constant	0.165***	-1.930***	-5.442***	-4.866***	-3.015***	-4.762***	-5.612***	-4.946***	-4.527***	-1.672***
	(0.018)	(0.140)	(0.286)	(0.321)	(0.261)	(0.448)	(0.421)	(0.399)	(0.278)	(0.138)
Observations	57134	57134	57134	57134	57134	57134	57134	57134	57134	57133

Standard errors in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

Bonferroni correction for 14 tests (in bold), or 7 tests (in italic) for all the regressors.

Table 3: Model 1: linear regression with robust standard errors clustered at the participant and post level. From Models 2 to 9: logit regression with robust standard errors clustered at the participant and post level.

C Overall Engagement: robustness check for topics

In Table 4, we check whether the posts' topics moderate the effects of the interaction between the conditions and the harmfulness of the topics for the overall engagement. We focus here only on the results found in the main analysis (see Table 3) - the interaction between the Cool and Descriptive conditions with the harmfulness of the posts and the topics. Assisted suicide and gun control topics moderate the effect of the interactions between the Cool and Descriptive conditions and the harmfulness of the posts. Specifically, the higher the harmfulness of the post the lower the overall engagement with the post. Moreover, legalisation and politics moderate the effect of the interaction between the Descriptive condition and the harmfulness of the posts, and in this case, the higher the harmfulness of the post the higher the overall engagement with the post.

	Abort	Ass. suicide	Gun cont	Legal	Politics	Science	Soc justice
Harmful	-0.005*	-0.005*	-0.005*	-0.003	-0.004	-0.006**	-0.007*
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)
Personal	0.064**	0.058**	0.066**	0.057**	0.061**	0.058**	0.066**
	(0.020)	(0.020)	(0.020)	(0.019)	(0.020)	(0.021)	(0.020)
Cool	0.053**	0.049*	0.054**	0.040*	0.047*	0.048*	0.048*
	(0.020)	(0.020)	(0.020)	(0.019)	(0.020)	(0.021)	(0.022)
Reputation	0.026	0.018	0.023	0.018	0.018	0.008	0.019
	(0.020)	(0.019)	(0.019)	(0.019)	(0.019)	(0.020)	(0.021)
Descriptive	0.125***	0.109***	0.119***	0.109***	0.116***	0.112***	0.118***
	(0.022)	(0.021)	(0.022)	(0.021)	(0.021)	(0.022)	(0.026)
Injunctive	0.084***	0.075***	0.084***	0.083***	0.078***	0.068**	0.085***
	(0.021)	(0.020)	(0.020)	(0.020)	(0.020)	(0.021)	(0.021)
Empathy	0.048*	0.040	0.048*	0.039*	0.042*	0.039	0.046*
	(0.020)	(0.020)	(0.020)	(0.019)	(0.020)	(0.021)	(0.022)
Deliberation	0.045*	0.038	0.047*	0.040*	0.040*	0.040	0.043*
	(0.020)	(0.020)	(0.020)	(0.019)	(0.020)	(0.021)	(0.020)
Personal×Harmful	-0.006*	-0.005	-0.006*	-0.005	-0.005	-0.004	-0.010*
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)
Cool×Harmful	-0.009**	-0.009**	-0.010***	-0.008**	-0.010***	-0.009**	-0.010*

	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)
Reputation×Harmful	-0.005	-0.003	-0.004	-0.003	-0.003	-0.002	-0.006
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)
Descriptive×Harmful	-0.013***	-0.011***	-0.012***	-0.011***	-0.013***	-0.011***	-0.015**
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)
Injunctive×Harmful	-0.008**	-0.006*	-0.008*	-0.008**	-0.006*	-0.005	-0.012**
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)
Empathy×Harmful	-0.003	-0.001	-0.002	-0.001	-0.002	-0.000	-0.006
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)
Deliberation×Harmful	-0.006*	-0.005	-0.007*	-0.006	-0.006	-0.006	-0.008*
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)
Topic	-0.025*	-0.022	0.022	0.070***	-0.044***	-0.026*	0.037*
	(0.011)	(0.021)	(0.019)	(0.018)	(0.010)	(0.011)	(0.016)
Personal×Topic	-0.044	0.371**	0.046	-0.023	-0.151*	0.021	-0.002
	(0.027)	(0.123)	(0.047)	(0.047)	(0.061)	(0.026)	(0.025)
Cool×Topic	-0.054*	0.296**	0.036	-0.039	-0.015	-0.008	-0.000
	(0.024)	(0.113)	(0.043)	(0.048)	(0.053)	(0.023)	(0.023)
Reputation×Topic	-0.080***	0.149	0.094	-0.044	-0.080	0.045	0.006
	(0.024)	(0.111)	(0.050)	(0.046)	(0.058)	(0.026)	(0.025)
Descriptive×Topic	-0.100***	0.549***	0.134*	-0.142**	-0.194**	0.007	0.002
	(0.026)	(0.138)	(0.054)	(0.049)	(0.061)	(0.027)	(0.037)
Injunctive×Topic	-0.057*	0.432***	0.088*	-0.040	-0.109*	0.052*	-0.010
	(0.025)	(0.121)	(0.044)	(0.047)	(0.052)	(0.025)	(0.025)
Empathy×Topic	-0.065*	0.405**	0.042	-0.021	-0.125	0.040	0.004
	(0.026)	(0.126)	(0.047)	(0.047)	(0.067)	(0.025)	(0.024)
Deliberation×Topic	-0.047	0.444**	0.086	-0.034	-0.002	-0.006	0.002
	(0.026)	(0.136)	(0.047)	(0.043)	(0.048)	(0.024)	(0.034)
Personal×Harmful×Topic	0.012*	-0.532**	-0.100**	0.040	0.024*	-0.010	0.006
	(0.006)	(0.169)	(0.033)	(0.045)	(0.010)	(0.006)	(0.005)

Cool×Harmful×Topic	0.008	**-0.477****	**-0.092****	0.090	0.007	0.002	0.002
	(0.004)	**(0.154)**	**(0.030)**	(0.047)	(0.009)	(0.006)	(0.006)
Reputation×Harmful×Topic	0.014**	-0.241	-0.151***	0.048	0.014	-0.009	0.003
	(0.005)	(0.153)	(0.035)	(0.043)	(0.010)	(0.006)	(0.005)
Descriptive×Harmful×Topic	0.012*	**-0.762*****	**-0.178*****	**0.215*****	**0.036****	-0.001	0.004
	(0.005)	**(0.189)**	**(0.041)**	**(0.055)**	**(0.011)**	(0.007)	(0.006)
Injunctive×Harmful×Topic	0.012*	-0.610***	-0.142***	0.015	0.016	-0.011	0.008
	(0.006)	(0.166)	(0.032)	(0.045)	(0.009)	(0.006)	(0.005)
Empathy×Harmful×Topic	0.014**	-0.588***	-0.098**	0.042	0.023*	-0.018**	0.006
	(0.005)	(0.174)	(0.032)	(0.046)	(0.012)	(0.006)	(0.005)
Deliberation×Harmful×Topic	0.007	-0.635***	-0.146***	0.038	0.001	0.004	0.003
	(0.005)	(0.182)	(0.032)	(0.041)	(0.008)	(0.006)	(0.006)
Constant	0.168***	0.167***	0.162***	0.151***	0.165***	0.173***	0.161***
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)	(0.017)
Observations	57134	57134	57134	57134	57134	57134	57134

Standard errors in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

Bonferroni correction for 7 tests in bold only for the interactions of interest.

Table 4: From Models 1 to 7: linear regressions with robust standard errors clustered at the participant and post levels.

D Like: robustness check for topics

In Table 5, we check whether the posts' topics moderate the effects of the interaction between the conditions and the harmfulness of the topics for the like intention. We focus here only on the results found in the main analysis (see Table 3) - the interaction between the Injunctive condition with the harmfulness of the post and the topic. Assisted suicide, gun control and legalization moderate the effect of the interaction between the Injunctive condition and the harmfulness of the post. Specifically. for assisted suicide and gun control, the higher the harmfulness of the post the lower the like intention to the post, while, for legalization the higher the harmfulness of the post the higher the like intention to the post.

	Abort	Ass. suicide	Gun cont	Legal	Politics	Science	Soc justice
Harmful	-0.265***	-0.274***	-0.275***	-0.254***	-0.283***	-0.273***	-0.282***
	(0.036)	(0.029)	(0.037)	(0.040)	(0.045)	(0.038)	(0.037)
Personal	0.458**	0.440**	0.506***	0.424**	0.467**	0.390*	0.399*
	(0.149)	(0.136)	(0.142)	(0.154)	(0.145)	(0.155)	(0.167)
Cool	0.377**	0.383**	0.393**	0.332**	0.411**	0.365*	0.294
	(0.131)	(0.136)	(0.133)	(0.123)	(0.132)	(0.142)	(0.151)
Reputation	0.294*	0.297*	0.309*	0.269	0.275*	0.228	0.233
	(0.135)	(0.141)	(0.136)	(0.144)	(0.135)	(0.128)	(0.157)
Descriptive	0.625***	0.580***	0.657***	0.551***	0.599***	0.592***	0.552***
	(0.137)	(0.132)	(0.130)	(0.132)	(0.136)	(0.151)	(0.159)
Injunctive	0.640***	0.624***	0.665***	0.703***	0.671***	0.559***	0.612***
	(0.140)	(0.134)	(0.143)	(0.149)	(0.142)	(0.136)	(0.156)
Empathy	0.333*	0.307*	0.372*	0.298	0.323*	0.229	0.304
	(0.160)	(0.139)	(0.154)	(0.180)	(0.160)	(0.158)	(0.186)
Deliberation	0.326*	0.319*	0.381**	0.258	0.357*	0.344*	0.327*
	(0.139)	(0.138)	(0.134)	(0.143)	(0.140)	(0.151)	(0.152)
Personal×Harmful	-0.144**	-0.121**	-0.128**	-0.119*	-0.141**	-0.111*	-0.080
	(0.051)	(0.044)	(0.048)	(0.048)	(0.050)	(0.054)	(0.066)
Cool×Harmful	-0.066	-0.067	-0.065	-0.061	-0.104*	-0.080	-0.001
	(0.038)	(0.038)	(0.037)	(0.035)	(0.041)	(0.044)	(0.052)
Reputation×Harmful	-0.076*	-0.069	-0.066	-0.066	-0.059	-0.067	-0.045

	(0.038)	(0.044)	(0.038)	(0.039)	(0.042)	(0.040)	(0.049)
Descriptive×Harmful	-0.076*	-0.070	-0.081*	-0.068*	-0.071	-0.095*	-0.049
	(0.035)	(0.039)	(0.035)	(0.034)	(0.039)	(0.042)	(0.046)
Injunctive×Harmful	-0.154***	-0.155***	-0.157***	-0.174***	-0.188***	-0.162***	-0.128*
	(0.040)	(0.039)	(0.040)	(0.041)	(0.048)	(0.048)	(0.054)
Empathy×Harmful	-0.070	-0.058	-0.068	-0.060	-0.071	-0.047	-0.041
	(0.045)	(0.042)	(0.044)	(0.047)	(0.049)	(0.046)	(0.061)
Deliberation×Harmful	-0.039	-0.033	-0.042	-0.025	-0.059	-0.047	-0.008
	(0.042)	(0.040)	(0.042)	(0.042)	(0.047)	(0.047)	(0.056)
Topic	-0.644	-0.051	-0.111	0.378*	0.211	-0.154	0.244
	(0.348)	(0.199)	(0.411)	(0.175)	(0.323)	(0.221)	(0.249)
Personal×Topic	-0.367	1.772*	0.205	-0.226	-2.064	0.245	0.083
	(0.281)	(0.788)	(0.381)	(0.439)	(1.537)	(0.254)	(0.301)
Cool×Topic	-0.390	2.727**	0.377	-0.425	-1.592	-0.049	0.149
	(0.320)	(0.930)	(0.377)	(0.427)	(1.776)	(0.234)	(0.248)
Reputation×Topic	-0.550*	1.898*	0.513	-0.463*	-2.078	0.185	0.105
	(0.267)	(0.962)	(0.383)	(0.214)	(1.102)	(0.339)	(0.221)
Descriptive×Topic	-0.855*	2.347**	0.139	-0.559	-1.488*	-0.040	0.111
	(0.388)	(0.806)	(0.367)	(0.419)	(0.690)	(0.230)	(0.245)
Injunctive×Topic	-0.307	2.251**	0.475	-0.641***	-1.931	0.352	0.021
	(0.447)	(0.735)	(0.370)	(0.181)	(2.004)	(0.233)	(0.264)
Empathy×Topic	-0.640	3.906***	0.089	-0.473	-0.681	0.383	-0.036
	(0.471)	(0.996)	(0.410)	(0.371)	(1.548)	(0.332)	(0.290)
Deliberation×Topic	-0.062	3.208***	0.278	-0.165	-1.128	-0.111	-0.075
	(0.418)	(0.911)	(0.384)	(0.250)	(1.446)	(0.233)	(0.247)
Personal×Harmful×Topic	0.246**	-2.710*	-1.123**	0.375	0.415	-0.054	-0.097
	(0.089)	(1.181)	(0.396)	(0.327)	(0.300)	(0.104)	(0.089)
Cool×Harmful×Topic	0.096	-4.546**	-0.839***	0.683*	0.380	0.090	-0.152
	(0.077)	(1.497)	(0.182)	(0.335)	(0.339)	(0.087)	(0.087)

Reputation×Harmful×Topic	0.199**	-3.298*	-1.295***	0.606***	0.338	0.016	-0.041
	(0.065)	(1.489)	(0.206)	(0.163)	(0.215)	(0.113)	(0.069)
Descriptive×Harmful×Topic	0.175	-3.475**	-0.928***	0.939**	0.251	0.112	-0.053
	(0.112)	(1.229)	(0.247)	(0.331)	(0.146)	(0.071)	(0.072)
Injunctive×Harmful×Topic	0.055	-3.342**	-1.080**	0.530**	0.417	0.000	-0.068
	(0.146)	(1.124)	(0.346)	(0.173)	(0.383)	(0.096)	(0.101)
Empathy×Harmful×Topic	0.206*	-6.234***	-0.890***	0.679*	0.153	-0.070	-0.030
	(0.089)	(1.623)	(0.114)	(0.268)	(0.312)	(0.108)	(0.081)
Deliberation×Harmful×Topic	0.042	-4.870***	-1.039***	0.497**	0.259	0.070	-0.051
	(0.103)	(1.432)	(0.077)	(0.158)	(0.278)	(0.069)	(0.076)
Constant	-1.874***	-1.923***	-1.913***	-2.028***	-1.921***	-1.893***	-1.976***
	(0.134)	(0.100)	(0.140)	(0.170)	(0.139)	(0.154)	(0.146)
Observations	57134	57134	57134	57134	57134	57134	57134

Standard errors in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

Bonferroni correction for 7 tests in bold only for the interactions of interest.

Table 5: Model 1: logit regression with robust standard errors clustered at the participant and post level. Model 2: logit regression with robust standard errors clustered at the participant level. From Models 3 to 7 logit regression with robust standard errors clustered at the participant and post level.

E Demographics: robustness check

In Table 6, we check whether gender and age moderate the effects of the interaction between the conditions and the harmfulness of the topics for both the overall engagement and the intention to like. We focus here only on the results found in the main analysis (see Table 3). For the overall engagement, we focus on the interaction between the Cool and Descriptive conditions with the harmfulness of the post and the demographic information. For the intention to like, we focus on the interaction between the Injunctive condition with the harmfulness of the post and the demographic information. For both the overall engagement and the intention to like, gender and age do not moderate the effect of the interactions between the conditions and the harmfulness of the post.

	Engagement	Like	Engagement	Like
	Female	Female	Age	Age
Harmful	-0.003	-0.219***	-0.003	-0.288**
	(0.003)	(0.042)	(0.008)	(0.093)
Personal	0.040	0.254	0.078	0.678
	(0.027)	(0.194)	(0.060)	(0.389)
Cool	0.031	0.380*	0.078	0.666
	(0.031)	(0.191)	(0.058)	(0.419)
Reputation	-0.006	0.205	0.036	0.365
	(0.029)	(0.189)	(0.061)	(0.463)
Descriptive	0.142***	0.815***	0.185**	0.731
	(0.034)	(0.181)	(0.062)	(0.382)
Injunctive	0.011	0.382*	0.088	0.564
	(0.030)	(0.193)	(0.061)	(0.414)
Empathy	-0.010	0.085	0.030	0.463
	(0.027)	(0.196)	(0.062)	(0.406)
Deliberation	0.023	0.282	0.007	0.380
	(0.033)	(0.219)	(0.063)	(0.445)
Personal×Harmful	-0.004	-0.096	-0.004	-0.195
	(0.004)	(0.059)	(0.010)	(0.142)
Cool×Harmful	-0.010*	-0.122*	-0.017*	-0.155

	(0.004)	(0.051)	(0.008)	(0.112)
Reputation×Harmful	-0.001	-0.053	0.001	0.064
	(0.004)	(0.051)	(0.010)	(0.158)
Descriptive×Harmful	-0.018***	-0.156**	-0.010	0.019
	(0.005)	(0.048)	(0.010)	(0.100)
Injunctive×Harmful	-0.003	-0.164**	-0.004	-0.138
	(0.004)	(0.054)	(0.009)	(0.116)
Empathy×Harmful	-0.000	-0.037	-0.008	-0.050
	(0.004)	(0.050)	(0.010)	(0.116)
Deliberation×Harmful	-0.009	-0.068	-0.008	-0.121
	(0.005)	(0.063)	(0.009)	(0.128)
Demographic	-0.019	-0.064	-0.000	0.001
	(0.030)	(0.240)	(0.001)	(0.009)
Harmful×Demographic	-0.003	-0.113	-0.000	0.000
	(0.005)	(0.060)	(0.000)	(0.002)
Personal×Demographic	0.047	0.483	-0.001	-0.006
	(0.039)	(0.263)	(0.002)	(0.010)
Cool×Demographic	0.039	0.059	-0.001	-0.008
	(0.045)	(0.293)	(0.002)	(0.011)
Reputation×Demographic	0.050	0.235	-0.001	-0.002
	(0.040)	(0.294)	(0.002)	(0.012)
Descriptive×Demographic	-0.052	-0.393	-0.002	-0.003
	(0.042)	(0.239)	(0.002)	(0.009)
Injunctive×Demographic	0.132**	0.512*	-0.000	0.002
	(0.040)	(0.255)	(0.002)	(0.010)
Empathy×Demographic	0.098*	0.491	0.000	-0.004
	(0.043)	(0.290)	(0.002)	(0.010)
Deliberation×Demographic	0.034	0.169	0.001	-0.001
	(0.042)	(0.288)	(0.002)	(0.011)

Personal×Harmful×Demographic	-0.004	-0.090	-0.000	0.002
	(0.005)	(0.082)	(0.000)	(0.003)
Cool×Harmful×Demographic	-0.001	0.101	0.000	0.002
	(0.006)	(0.075)	(0.000)	(0.003)
Reputation×Harmful×Demographic	-0.004	-0.030	-0.000	-0.004
	(0.006)	(0.083)	(0.000)	(0.004)
Descriptive×Harmful×Demographic	0.012*	0.176**	-0.000	-0.003
	(0.005)	(0.068)	(0.000)	(0.002)
Injunctive×Harmful×Demographic	-0.008	0.026	-0.000	-0.001
	(0.006)	(0.066)	(0.000)	(0.003)
Empathy×Harmful×Demographic	-0.003	-0.069	0.000	-0.000
	(0.006)	(0.076)	(0.000)	(0.003)
Deliberation×Harmful×Demographic	0.006	0.064	0.000	0.002
	(0.006)	(0.075)	(0.000)	(0.003)
Constant	0.178***	-1.889***	0.178***	-1.943***
	(0.022)	(0.169)	(0.043)	(0.341)
Observations	53197	53197	54877	54877

Standard errors in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

Bonferroni correction for 7 tests in bold only for the interactions of interest.

Table 6: Models 1 and 3: linear regressions with robust standard errors clustered at the participant and post level. Models 2 and 4: logit regressions with robust standard errors clustered at the participant and post level.

F Political orientation: robustness check

In Table 7, we check whether political orientation moderates the effects of the interaction between the conditions and the harmfulness of the topics for both the overall engagement and the like intention. We focus here only on the results found in the main analysis (see Table 3). For the overall engagement, we focus on the interaction between the Cool and Descriptive conditions with the harmfulness of the post and the demographic information. For the like intention, we focus on the interaction between the Injunctive condition with the harmfulness of the post and the demographic information. For both the overall engagement and the like intention, the political orientation does not moderate the effect of the interactions between the conditions and the harmfulness of the post.

	Engagement	Like	Engagement	Like	Engagement	Like
	Dems	Dems	Reps	Reps	Indeps	Indeps
Harmful	-0.002	-0.239***	-0.006	-0.309***	-0.004	-0.262***
	(0.003)	(0.048)	(0.003)	(0.047)	(0.003)	(0.040)
Personal	0.056*	0.266	0.052*	0.405**	0.075**	0.590***
	(0.025)	(0.218)	(0.022)	(0.145)	(0.023)	(0.162)
Cool	0.073**	0.506**	0.030	0.244	0.060*	0.489**
	(0.025)	(0.163)	(0.021)	(0.129)	(0.023)	(0.152)
Reputation	0.057*	0.479*	0.000	0.220	0.017	0.248
	(0.025)	(0.207)	(0.020)	(0.139)	(0.022)	(0.149)
Descriptive	0.143***	0.703***	0.105***	0.507***	0.091***	0.612***
	(0.030)	(0.189)	(0.025)	(0.137)	(0.025)	(0.162)
Injunctive	0.060*	0.593**	0.082***	0.603***	0.091***	0.728***
	(0.025)	(0.200)	(0.022)	(0.135)	(0.023)	(0.160)
Empathy	0.046*	0.247	0.034	0.289	0.052*	0.383*
	(0.023)	(0.201)	(0.022)	(0.156)	(0.024)	(0.172)
Deliberation	0.059*	0.398*	0.031	0.291*	0.040	0.330*
	(0.025)	(0.189)	(0.022)	(0.146)	(0.023)	(0.157)
Personal×Harmful	-0.007	-0.087	-0.006	-0.116*	-0.006	-0.144*
	(0.004)	(0.063)	(0.003)	(0.057)	(0.003)	(0.058)
Cool×Harmful	-0.009*	-0.071	-0.007*	-0.032	-0.011**	-0.080*

	(0.004)	(0.049)	(0.003)	(0.039)	(0.003)	(0.040)
Reputation×Harmful	-0.002	-0.056	-0.002	-0.077	-0.006	-0.073
	(0.004)	(0.061)	(0.003)	(0.043)	(0.003)	(0.042)
Descriptive×Harmful	-0.008*	-0.057	-0.012**	-0.050	-0.013***	-0.106**
	(0.004)	(0.051)	(0.004)	(0.040)	(0.003)	(0.040)
Injunctive×Harmful	-0.010*	-0.163**	-0.008*	-0.138***	-0.007*	-0.157***
	(0.004)	(0.061)	(0.003)	(0.041)	(0.003)	(0.043)
Empathy×Harmful	-0.001	0.013	-0.002	-0.084	-0.002	-0.076
	(0.003)	(0.057)	(0.003)	(0.049)	(0.004)	(0.052)
Deliberation×Harmful	-0.004	-0.004	-0.006	-0.045	-0.007*	-0.030
	(0.003)	(0.057)	(0.003)	(0.050)	(0.003)	(0.049)
Political Orientation	0.094**	0.557**	-0.113***	-1.371***	-0.000	0.124
	(0.030)	(0.198)	(0.031)	(0.305)	(0.029)	(0.218)
Harmful×PO	-0.005	-0.059	0.007	0.314***	-0.005	-0.045
	(0.004)	(0.069)	(0.005)	(0.071)	(0.004)	(0.078)
Personal×PO	0.016	0.360	0.066	0.745*	-0.063	-0.651*
	(0.044)	(0.294)	(0.048)	(0.329)	(0.042)	(0.300)
Cool×PO	-0.051	-0.218	0.134**	1.328***	-0.048	-0.484
	(0.039)	(0.247)	(0.047)	(0.372)	(0.040)	(0.259)
Reputation×PO	-0.081*	-0.350	0.129*	0.995*	0.001	0.108
	(0.037)	(0.279)	(0.049)	(0.393)	(0.043)	(0.304)
Descriptive×PO	-0.058	-0.152	0.077	1.094**	0.082	-0.063
	(0.043)	(0.236)	(0.051)	(0.348)	(0.049)	(0.303)
Injunctive×PO	0.043	0.106	-0.011	0.648	-0.052	-0.356
	(0.040)	(0.267)	(0.041)	(0.337)	(0.043)	(0.296)
Empathy×PO	-0.011	0.144	0.047	0.578	-0.036	-0.273
	(0.038)	(0.232)	(0.049)	(0.415)	(0.040)	(0.258)
Deliberation×PO	-0.033	-0.042	0.077	0.854*	-0.001	-0.003
	(0.038)	(0.244)	(0.048)	(0.356)	(0.043)	(0.275)

Personal×Harmful×PO	0.003	-0.070	0.000	-0.126	0.002	0.099
	(0.006)	(0.098)	(0.007)	(0.081)	(0.006)	(0.126)
Cool×Harmful×PO	0.001	0.003	-0.011	-0.287***	0.008	0.066
	(0.005)	(0.065)	(0.006)	(0.077)	(0.005)	(0.080)
Reputation×Harmful×PO	-0.004	-0.036	-0.007	-0.097	0.013	0.036
	(0.005)	(0.082)	(0.006)	(0.093)	(0.007)	(0.104)
Descriptive×Harmful×PO	-0.007	-0.048	-0.001	-0.227**	0.009	0.116
	(0.005)	(0.065)	(0.007)	(0.086)	(0.006)	(0.093)
Injunctive×Harmful×PO	0.005	0.003	0.002	-0.197*	-0.000	-0.033
	(0.006)	(0.083)	(0.006)	(0.098)	(0.006)	(0.105)
Empathy×Harmful×PO	-0.002	-0.163	0.007	0.030	0.002	0.066
	(0.006)	(0.086)	(0.007)	(0.083)	(0.006)	(0.101)
Deliberation×Harmful×PO	-0.004	-0.094	-0.002	-0.087	0.006	-0.024
	(0.005)	(0.085)	(0.006)	(0.080)	(0.006)	(0.088)
Constant	0.119***	-2.230***	0.179***	-1.800***	0.165***	-1.960***
	(0.016)	(0.154)	(0.020)	(0.144)	(0.021)	(0.162)
Observations	57134	57134	57134	57134	57134	57134

Standard errors in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

Bonferroni correction for 7 tests in bold only for the interactions of interest.

Table 7: Models 1, 3 and 5: linear regressions with robust standard errors clustered at the participant and post level. Models 2, 4 and 6: logit regressions with robust standard errors clustered at the participant and post level.

G Text Analysis

G.1 Definitions of the metrics

The following definitions are taken from the PeRspective API website.

Toxicity. A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.

Severe toxicity. A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words.

Identity attack. Insulting, inflammatory, or negative comment towards a person or a group of people.

Profanity. Swear words, curse words, or other obscene or profane language.

Threat. Describes an intention to inflict pain, injury, or violence against an individual or group.

Sexually explicit. Contains references to sexual acts, body parts, or other lewd content.

Flirtation. Pickup lines, complimenting appearance, subtle sexual innuendos, etc.

Attack on the author. Attack on the author of an article or post.

Attack on commenter. Attack on fellow commenter.

Incoherent. Difficult to understand, nonsensical.

Inflammatory. Intending to provoke or inflame.

Likely to reject. Overall measure of the likelihood for the comment to be rejected according to the NYT's moderation.

Obscene. Obscene or vulgar language such as cursing.

Spam. Irrelevant and unsolicited commercial content.

Unsubstantial. Trivial or short comments.

G.2 Analysis

We conducted a series of exploratory analyses to test for differences in the style of the comments left in each condition (Tables 8 to 10). Models that detect the harmful intent of comments (e.g., identity attacks, insults, threats) show no significant differences or weakly significant differences ($p > 0.01$) that do not survive correction for multiple comparisons. However, two metrics show a more consistent pattern: comment substance and coherence. Several interventions show a significant interaction between post harmfulness and experimental condition: compared to baseline, substance and coherence tend to increase as the original content becomes more harmful. For both metrics, one intervention survives multiple comparisons: highlighting reputation. In Table 8 we report the regression analysis regarding these two metrics. The regression tables of the other metrics are reported in Tables 9 and 10.

	Unsubstantial	Incoherent
Harmful	0.108*	0.058
	(0.049)	(0.062)
Personal	0.097	0.079
	(0.053)	(0.051)
Cool	0.140*	0.043
	(0.056)	(0.040)
Reputation	0.126*	0.152*
	(0.060)	(0.059)
Descriptive	0.147**	0.108
	(0.049)	(0.056)
Injunctive	0.088	0.056
	(0.052)	(0.046)
Empathy	0.098	0.112*
	(0.054)	(0.048)
Deliberation	0.124*	0.089*
	(0.053)	(0.043)
Personal×Harmful	-0.184*	-0.134
	(0.074)	(0.093)
Cool×Harmful	-0.178*	-0.028
	(0.068)	(0.069)

Reputation×Harmful	-0.249***	-0.298***
	(0.065)	(0.080)
Descriptive×Harmful	-0.205*	-0.271**
	(0.077)	(0.096)
Injunctive×Harmful	-0.131	-0.137
	(0.078)	(0.085)
Empathy×Harmful	-0.156*	-0.230*
	(0.076)	(0.088)
Deliberation×Harmful	-0.191**	-0.202**
	(0.071)	(0.073)
Constant	0.635***	0.379***
	(0.040)	(0.039)
Observations	1560	1560

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Linear regressions with robust standard errors clustered at the participant and post level.

	Severe Toxicity	Identity Attack	Insult	Profanity	Sexually Explicit	Threat	Flirtation
Harmful	0.063**	0.081***	0.246***	0.192**	0.051***	0.014	0.065*
	(0.021)	(0.017)	(0.054)	(0.057)	(0.014)	(0.024)	(0.029)
Personal	-0.006	0.030*	-0.012	-0.003	0.013	0.024	-0.005
	(0.012)	(0.012)	(0.033)	(0.030)	(0.011)	(0.014)	(0.024)
Cool	-0.015	-0.006	-0.009	-0.016	-0.012	0.014	0.005
	(0.012)	(0.021)	(0.048)	(0.032)	(0.013)	(0.021)	(0.027)
Reputation	-0.026	-0.005	-0.017	-0.016	-0.012	-0.015	0.038
	(0.015)	(0.016)	(0.042)	(0.031)	(0.009)	(0.011)	(0.031)
Descriptive	-0.016	-0.006	-0.036	-0.024	-0.012	-0.006	0.007
	(0.012)	(0.015)	(0.024)	(0.029)	(0.008)	(0.011)	(0.022)
Injunctive	-0.004	-0.000	0.017	0.024	0.004	0.001	0.027
	(0.011)	(0.018)	(0.032)	(0.031)	(0.009)	(0.010)	(0.028)
Empathy	-0.020	-0.002	-0.039	-0.026	-0.006	-0.002	0.004
	(0.013)	(0.017)	(0.035)	(0.033)	(0.012)	(0.011)	(0.018)
Deliberation	0.002	0.005	0.025	0.028	0.007	0.028*	0.008
	(0.013)	(0.015)	(0.036)	(0.033)	(0.011)	(0.012)	(0.024)
Personal×Harmful	0.002	-0.027	0.015	-0.018	-0.018	-0.041	-0.002
	(0.029)	(0.020)	(0.073)	(0.062)	(0.025)	(0.028)	(0.042)
Cool×Harmful	0.025	0.039	-0.007	0.004	0.040	-0.004	-0.018
	(0.025)	(0.037)	(0.100)	(0.069)	(0.030)	(0.039)	(0.046)
Reputation×Harmful	0.066	0.018	0.065	0.079	0.050*	0.018	-0.077
	(0.038)	(0.028)	(0.086)	(0.067)	(0.022)	(0.031)	(0.056)
Descriptive×Harmful	0.028	-0.003	0.077	0.063	0.025	0.007	-0.033
	(0.028)	(0.033)	(0.065)	(0.075)	(0.022)	(0.026)	(0.036)
Injunctive×Harmful	0.023	0.024	0.002	-0.039	0.004	0.016	-0.067
	(0.030)	(0.038)	(0.075)	(0.076)	(0.022)	(0.027)	(0.041)
Empathy×Harmful	0.032	0.035	0.055	0.030	0.016	-0.001	0.001
	(0.028)	(0.027)	(0.084)	(0.069)	(0.024)	(0.024)	(0.033)

Deliberation×Harmful	-0.005	0.011	-0.012	-0.067	-0.009	-0.041	-0.050
	(0.028)	(0.019)	(0.071)	(0.063)	(0.018)	(0.028)	(0.038)
Constant	0.003	0.012	0.042	0.030	0.016*	0.022*	0.245***
	(0.011)	(0.011)	(0.026)	(0.026)	(0.007)	(0.010)	(0.018)
Observations	1560	1560	1560	1560	1560	1560	1560

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 9: Linear regressions with robust standard errors clustered at the participant and post level.

	Attack on Aut	Attack on Comm	Toxicity	Inflammatory	Likely to Reject	Obscene	Spam
Harmful	0.256***	0.425***	0.331***	0.166**	0.401***	0.295***	0.001
	(0.019)	(0.095)	(0.073)	(0.050)	(0.074)	(0.075)	(0.024)
Personal	0.023	-0.021	0.009	0.047	0.040	0.040	-0.008
	(0.034)	(0.061)	(0.041)	(0.050)	(0.062)	(0.034)	(0.026)
Cool	0.031	0.031	-0.005	0.001	0.120*	0.013	0.003
	(0.027)	(0.068)	(0.051)	(0.054)	(0.058)	(0.038)	(0.022)
Reputation	0.045	0.037	-0.017	0.059	0.110	0.075	0.026
	(0.036)	(0.070)	(0.047)	(0.053)	(0.070)	(0.041)	(0.015)
Descriptive	0.025	-0.095	-0.048	-0.032	0.029	0.013	0.051
	(0.030)	(0.055)	(0.034)	(0.038)	(0.059)	(0.034)	(0.047)
Injunctive	0.008	-0.030	0.019	0.034	0.079	0.067	0.008
	(0.024)	(0.061)	(0.038)	(0.049)	(0.061)	(0.039)	(0.022)
Empathy	-0.005	-0.034	-0.051	0.014	0.003	0.045	0.022
	(0.023)	(0.054)	(0.043)	(0.046)	(0.068)	(0.041)	(0.026)
Deliberation	0.001	-0.071	0.035	0.060	0.082	0.087	-0.011
	(0.031)	(0.056)	(0.043)	(0.039)	(0.060)	(0.047)	(0.024)
Personal×Harm	-0.099	-0.046	-0.056	-0.068	-0.098	-0.137	0.038
	(0.062)	(0.101)	(0.084)	(0.070)	(0.088)	(0.080)	(0.045)
Cool×Harm	-0.105*	-0.098	-0.025	-0.012	-0.152	-0.046	0.002
	(0.045)	(0.118)	(0.100)	(0.080)	(0.079)	(0.102)	(0.035)
Reputation×Harm	-0.097	-0.022	0.045	-0.085	-0.068	-0.048	-0.042
	(0.072)	(0.102)	(0.095)	(0.085)	(0.095)	(0.080)	(0.023)
Descriptive×Harm	-0.048	0.050	0.069	-0.038	0.055	-0.036	-0.094
	(0.041)	(0.103)	(0.083)	(0.061)	(0.081)	(0.092)	(0.072)
Injunctive×Harm	-0.005	0.063	-0.026	-0.034	-0.058	-0.170	-0.034
	(0.039)	(0.112)	(0.090)	(0.078)	(0.083)	(0.109)	(0.028)
Empathy×Harm	-0.005	0.006	0.057	-0.053	0.000	-0.117	-0.055
	(0.041)	(0.092)	(0.098)	(0.074)	(0.090)	(0.087)	(0.036)

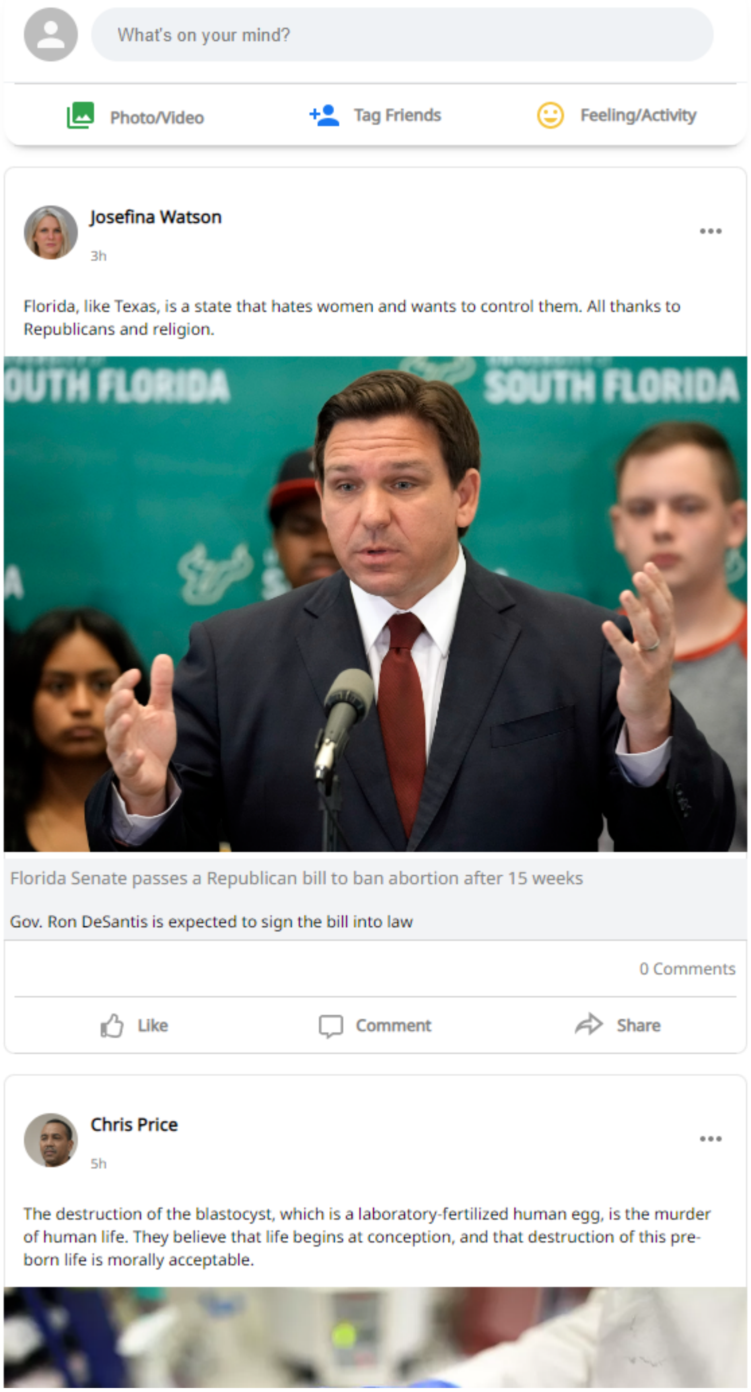
Deliberation×Harm	-0.075	0.067	-0.064	-0.109*	-0.123	-0.223*	-0.001
	(0.072)	(0.087)	(0.084)	(0.042)	(0.084)	(0.085)	(0.033)
Constant	0.068***	0.270***	0.077*	0.234***	0.474***	0.020	0.070***
	(0.017)	(0.053)	(0.035)	(0.036)	(0.056)	(0.029)	(0.020)
Observations	1560	1560	1560	1560	1560	1560	1560

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 10: Linear regressions with robust standard errors clustered at the participant and post level. Attack on Aut = attack on author of the post; Attack on Comm = attack on another commenter of the post

H Sample of the Facebook's newsfeed used in the experiment



Josefina Watson

3h

Florida, like Texas, is a state that hates women and wants to control them. All thanks to Republicans and religion.

Florida Senate passes a Republican bill to ban abortion after 15 weeks

Gov. Ron DeSantis is expected to sign the bill into law

Chris Price

5h

The destruction of the blastocyst, which is a laboratory-fertilized human egg, is the murder of human life. They believe that life begins at conception, and that destruction of this preborn life is morally acceptable.

Figure 2: Sample of the Facebook's newsfeed used in the experiment