Public's demand for moderation to online toxic speech is limited and influenced by social factors*

Yichen Ji Xiaoxu Liu

February 15, 2024

In this report,we analysis the public's ideal moderation for 5 levels toxic online speech.It was found that most people tend to not give harsh moderations, unless the speech touches upon personal threat. And people with different social factors have difference preferences on moderations.The findings will help me build a ideal standard platform for the public.

Table of contents

*Code and data are available at: https://github.com/Selinayichenji/Toxic_speech_replication.git.

1 Introduction

*This reproduction was performed after a replication on the Social Science Reproduction platform:**link here***

The research has shown that roughly 4 in 10 Americans have experienced online harassment, including name-calling, physical threats, and sexual abuse (Vogels 2021). With the current widespread use of social media, the debation of regulating toxic content becomes even more pertinent. Cultivating civility within democratic discourse is strongly necessary, such as emphasizing respect and social order when communicating online. However, the emergence of uncivil, intolerant content on social platforms raises concerns about its potential harm to public discourse and democracy. Our focus is on the crucial dilemma: should measures be implemented to moderate toxic content and uphold civility, or should we suggest allowing such speech on social media to remain unconstrained? The pursuit of addressing these challenges becomes significant in shaping the future of online democratic engagement.

Although hate, harassment, and extremism significantly impact the country and online community negatively as toxic comments have saturated lots of common social media in the U.S., the application of strong and effective regulation over social media faces challenges due to multiple factors. Various factors, including technology companies, government, and NGOs, oppose the potential heavy regulation, which operates within a distinct legal framework in the U.S. Besides, Users, the ultimate recipients of online toxicity, are important in reporting objectionable content through flagging mechanisms. Therefore, we attempt to find the ideal platform standards from users' views to against toxic speech.

We aim to apply the initial analysis from the original paper "Toxic Speech and Limited Demand for Content Moderation on Social Media", which is from the American Political Science Review. The paper attempts to figure out the consistency of how users reply to toxic speech when using social media to find an appropriate solution to improve the harmony of online platforms, and it includes two pieces of research: 1. targeting social groups and 2. targeting partisans.

In our reproduction, we use the original methods and the same dataset and shift the concentration from the level of toxic speech to the types of users while extending the research with more specific prevalent aspects of the groups of people's responses. Instead of talking about how people respond to toxic speech toward different labels of victims (LGBTQ, billionaire, and Christian) in study 1 of the original paper, we focused on how people with different genders, education levels, and races react to toxic speech toward LGBTQ. The estimand of our paper is differences of tendency of giving content moderation among people with different gender, race and education level when they face the toxic speech's target is LGBTQ.Beyond the changes, we hold all other perspectives to be the same as the original paper.

We obtain the result of the reproduction work and find that white people, people have post-graduate degree and man tend to give slighter moderation than other groups. Woman and black people tend to give serious moderation than others. The research result provides us with the information that: - In the traditional sense, socially advantaged groups tend to propose more lenient content moderation measures when the LGBTQ community is attacked, whereas disadvantaged groups are inclined to penalize speech that attacks the LGBTQ community. - People usually want a more loose-controlled online environment and choose no heavy moderation toward heavy speech.

These help us understand the user preference in social media, and stimulate the appropriate regulations of the speeches on online platforms. Generally, the reproduction talks about the summary of what we do based on the original paper and what we obtain, the data sources, detailed pictures and analysis through coding, and the discussion that concludes our results and lessons while discussing (potential) drawbacks and anticipated regulations/behaviors in the future.

The paper introduces the basic information contains data source, methodology, variables and measurements in Section 2. The visualization and analysis of the tendency differences in moderation by people across three social factors are presented in Section 3. And in Section 4, we discuss what we have learned, our understanding of the world, the limitations, and the next steps of our research. Section A is for appendix and Section A is listed all references in this paper.

2 Data

2.1 Source

Our replication paper is based on the original paper in American Political Science Review "Toxic Speech and Limited Demand for Content Moderation on Social Media". Our paper is consistent with the original goal that attempts to find how people respond to toxic speech with different targets to decide the moderation of social media for a respectful and harmonious environment. The dataset and replication package is open on the Harvard Dataverse website. For the data in LGBTQ target, the only esstential one is lgbtq.RData in the replication package.

The data came from the survey results collected by the authors of original paper. So there is no other similar datasets for same research.

2.2 Methodology

Our paper applies using the statistical programming language R (R Core Team 2022). Besides the programming tool, we also employ the following packages: readr (Wickham, Hester, and Bryan 2023), ggplot2 (Wickham 2016), dplyr (Wickham et al. 2023), tidyverse (Wickham et al. 2019), MetBrewer (Mills 2022), knitr (Xie 2023), tidyr(Wickham, Vaughan, and Girlich 2023), kableExtra(Zhu 2021) and grid(R Core Team 2023).

2.3 Variables

We only introduced the variables used in our own analysis. For the full variables in survey from the original paper, please check the Appendix.

- Treatment: non-group-related control, control, uncivil, intolerant and threatening.
  - Non-group-related control means no target and no toxic language.
  - Control means anti-target but without the 3 kinds of toxic languages: uncivil, intolerant and threatening.
  - Uncivil was defined as "including anything from an unnecessarily disrespectful tone and lack of respect to rudeness and inconsiderate language."(PRADEL F, n.d.).
  - The intolerance differs from incivility, "it aims to derogate, silence, or undermine particular groups due to their protected characteristics, attack their rights, and incite violence and harm."(PRADEL F, n.d.).
  - The threatening is a toxic behavior explicitly announces the intention of physical harm.(PRADEL F, n.d.).
- Handle: (1) Leave it, do nothing; (2) Place a warning label on the post; (3) Reduce how many people can see the post; (4) Permanently remove the post; (5) Suspend the person's account.
- Gender: Male, Female and Others.
- Education: High school graduate; College; Postgraduate.
- Race: White, Black, Hispanic and Others.
- Percentage: It means the ratio of 5 handles given by people with the same social factor when they are facing the same treatment.

2.4 Data examples

Table 1: Example of Analysis Data in LGBTQ Topic

treatment	handle	gender	race	educ
uncivil	Leave it, do nothing	Female	Black	College
control	Leave it, do nothing	Female	White	High school graduate
non-group-related	Leave it, do nothing	Female	White	College
non-group-related	Leave it, do nothing	Male	White	Postgraduate
uncivil	Permanently remove the post	Female	White	Postgraduate
threatening	Suspend the person's account	Female	White	High school graduate

2.5 Measurement

The data collected by using survey. In the study I (target social group), targets include LGBTQ, Billionaire and Christians. The study measures participants' preferences for handling toxic speech on social media by asking them to react to posts on social media platforms. Specifically, the research measures participants' preferences through the following question:

- "In your view, how should social media companies (such as Facebook and Twitter) handle the post above?" Participants can choose from the following options:
  - "Leave it, do nothing"
  - "Place a warning label on the post"
  - "Reduce how many people can see the post"
  - "Permanently remove the post"
  - "Suspend the person's account"

These options allow participants to express their preferences for the actions social media platforms should take regarding toxic speech. By observing participants' choices among these options, researchers can understand participants' varying preferences for content moderation on social media platforms under different experimental conditions.

The experiments took place in July 2022 (LGBTQ and Billionaires) and October 2022 (Christians). Each study recruited between 1,300 and 2,000 U.S. adults from the participant pool of the crowdsourcing platform Prolific. Participants were recruited according to specific procedures outlined in the APSR Dataverse (Pradel et al. 2024). Exclusion criteria, as pre-registered, included participants who failed the attention check, opted for exclusion from the study, or completed the survey in less than 50 seconds. Ethics approval was obtained from the Central University Research Ethics Committee of the University of Oxford, and the study design was pre-registered prior to data collection.(PRADEL F, n.d.)

3 Results

We selected 3 social factors (gender, education level and race) to investigate people's moderation preference in LGBTQ Target toxic speech.

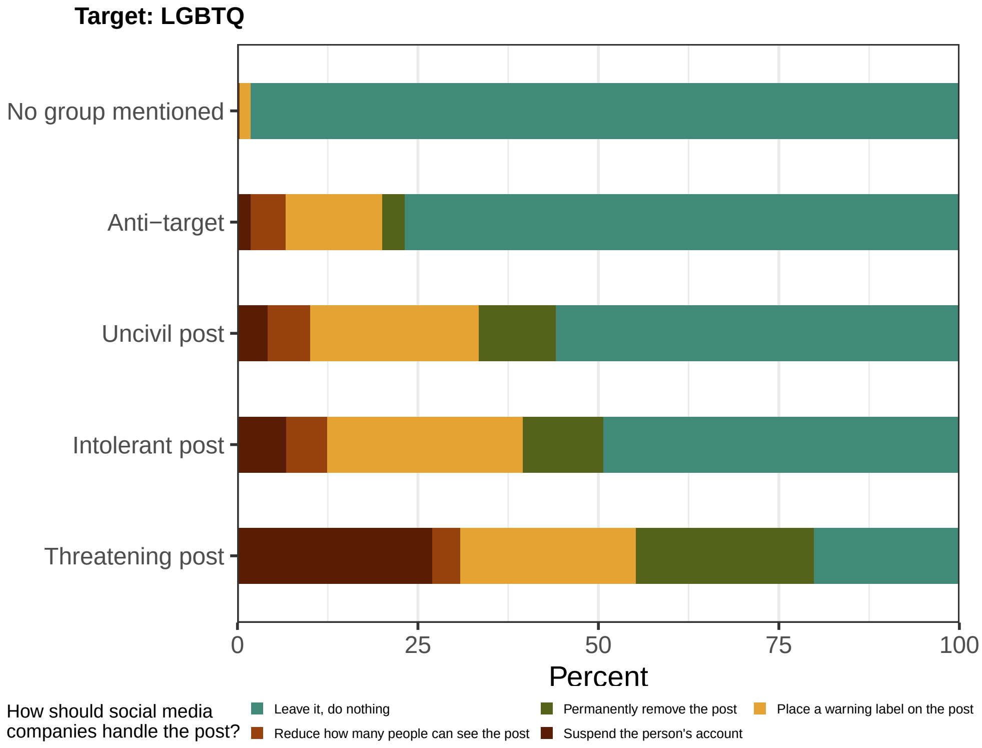

Figure 1: Preferred actions in response to distinct post types - LGBTQ

Figure 1 illustrates the percentages of 5 different handles in 5 levels toxic speech targets LGBTQ. As shown on the Figure 1, the serious content moderations (Suspend account and Permanently remove the post) take more percentage as the toxic level increases. The interviewees are especially sensitive to threatening post to LGBTQ people. We can see that for no group mentioned post, most people choose to leave it, a few people put a warning. But in Threatening post, the percentages of permanently remove the post and suspend the account increase obviously, Interestingly, the indirect moderations'(Reduce posts' view and Put a warning label) percentages in threatening posts decrease compared to Intolerant post, but percentages direct moderations(Suspend account and Permanently remove the post) increase. This means people give up indirect method and more tend to take though and serious moderation.

In Figure 2, Figure 3 and Figure 4, we visualize the difference of tendency of different groups' moderation. The plots arranges in order of the seriousness of moderation. Treatment is in the x-axis, ordered in posts' toxicity level, 1 to 5, from the lightest to the most toxic posts.

Percentages are in y-axis. The percentage means what percent of people in specific social group takes such handle/moderation when they see the posts in x-axis. For example, the sum of all woman's percentages in Treatment 1 is 100%. If you are interested in specific numeric numbers, please check Table 3 to 5 in appendix.

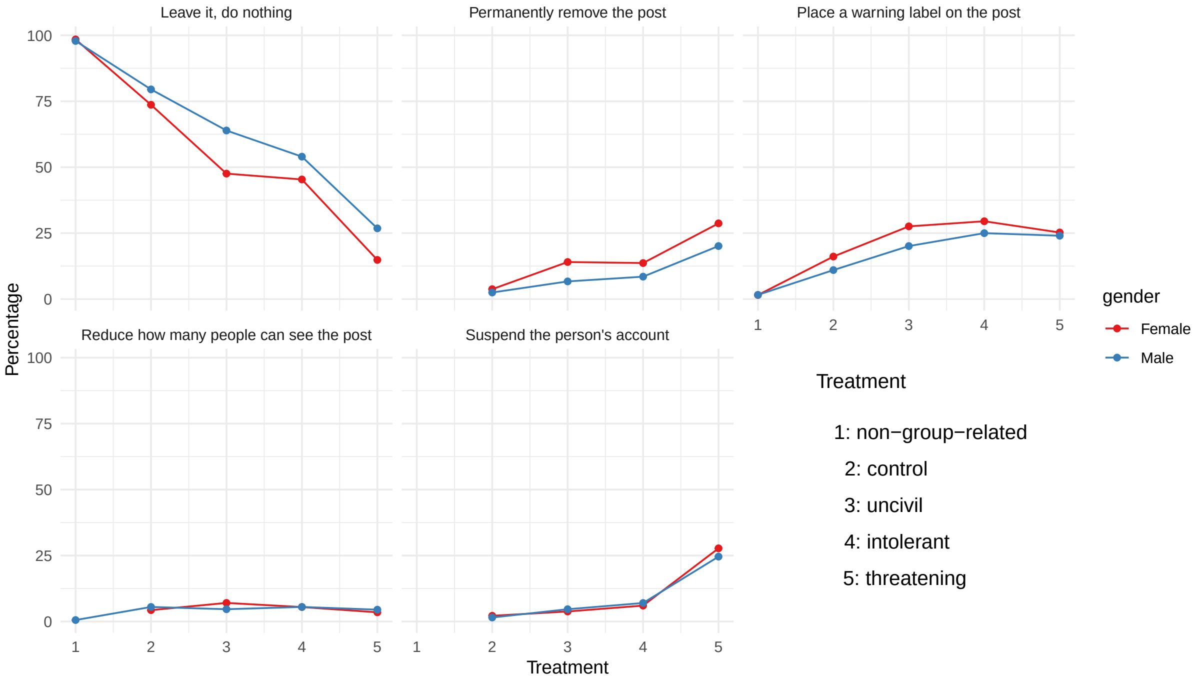

Figure 2: Percentage by Treatment and Gender in LGBTQ Topic

From Figure 2, female tends to give more harsh moderation than male. Males are more inclined to take no actions than female in all toxic levels.As the toxicity of the posts increases, the tendency of females to do nothing significantly decreases, especially for threatening posts. Both men and women are more inclined to permanently remove a post when it is threatening, but the tendency increases more sharply for women. As the toxicity of posts escalates from uncivil to threatening, both men and women are more likely to place a warning label, with women showing a slightly steeper increase in this tendency. There is no significant gender difference for reducing posts' views, with both males and females less likely to choose this moderation for all toxicity levels of posts. Females are more likely than males to opt for suspending the account when dealing with threatening posts. In contrast, males maintain a relatively stable tendency to moderate less than woman across all levels of post toxicity.

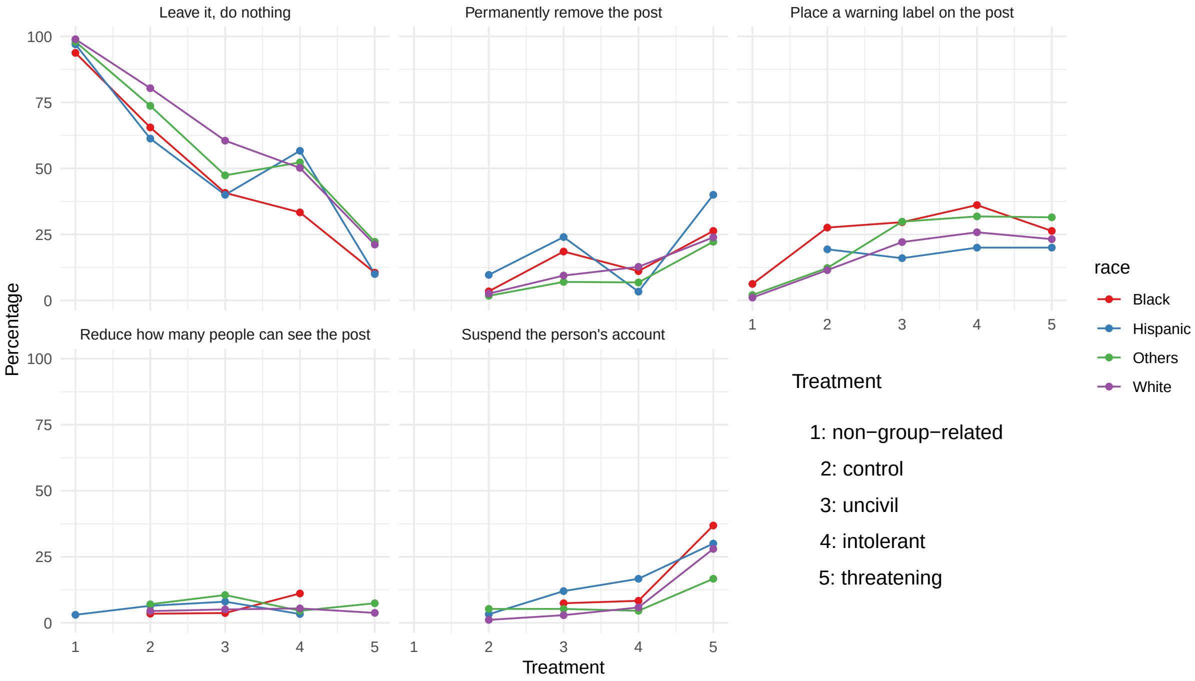

Figure 3: Percentage by Treatment and Race in LGBTQ Topic

As shown on Figure 3, across all levels of post toxicity, Black individuals have the lowest tendency to take no action, especially noticeable in non-group-related (1) and control (2) posts, where their proportion of inaction rapidly decreases. Whites start with the highest proportion of inaction, which also decreases as the toxicity of the posts increases. Hispanics and Others have a moderate tendency not to act, which decreases with rising toxicity. When facing threatening posts (5), all racial groups show an increased tendency to remove the post, with Blacks and Hispanics showing the most significant upward trend. As the toxicity of posts increases, all racial groups are more inclined to place warning labels on posts, with Whites and Others showing a more noticeable trend upwards. There is little difference among racial groups in terms of reducing a post's visibility, with a general tendency not to choose this action, especially when dealing with less toxic posts. When faced with threatening posts, the propensity to suspend accounts increases among all racial groups, with Blacks and Hispanics showing a particularly significant upward trend.

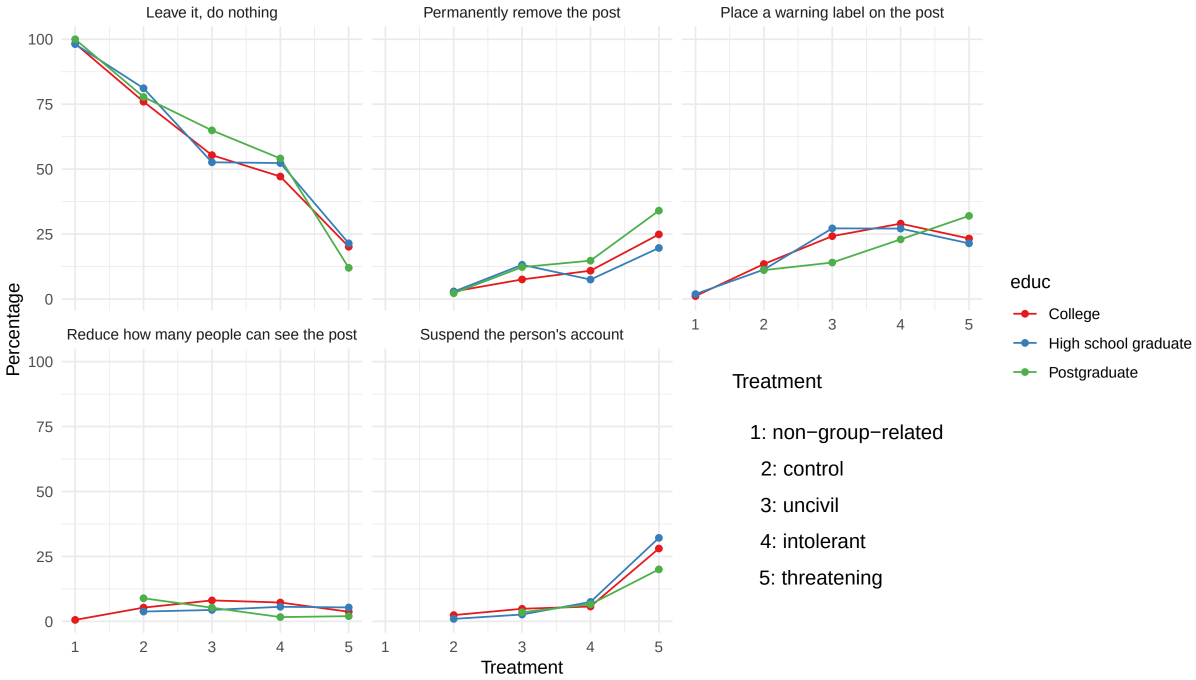

Figure 4: Percentage by Treatment and Education in LGBTQ Topic

As shown on Figure 4, the inclination to take no action across all toxicity levels is relatively similar among the three education levels, with a steady decline as the posts increase in toxicity.

At toxicity levels 1 to 2, there is no significant difference among the three educational backgrounds regarding the percentage who prefer not to take any action. At toxicity levels 3 to 4, the percentage of postgraduates who choose not to take any action is noticeably higher than the other two educational backgrounds. However, when the toxicity level is at its most severe, it is the postgraduates who have the smallest proportion unwilling to take action. Comparing the other four graphs, it can be seen that at toxicity levels 3 to 4, postgraduates have a lower percentage only in terms of putting a warning label on content. For the other three moderation actions, there is no significant difference compared to the other two educational backgrounds. However, when the toxicity level reaches the threshold of being threatening, the percentages for permanently removing the post and placing a warning label are the highest among the three educational backgrounds. Interestingly, for the most severe measure—account suspension—the percentage is actually the lowest.

In conclusion, the chart suggests that as posts become more toxic, individuals across all educational backgrounds tend to adopt stricter moderation actions. However, the variations among different educational levels are modest, with postgraduates slightly more inclined towards slighter measures (compare to suspend the account) such as putting warning label and remove the posts when dealing with the most toxic posts.

4 Discussion

4.1 What we did

We replicated 3 figures contains 2 panels “Target LGBTQ” and “Target: Billionaires” in Figure3 “Preferences for Content Moderation by Treatment and by Experiment in Study I” and 1 element, Figure 6 “Preferences for Specific Types of Content Moderation by Treatment in Study II” from the paper “Toxic Speech and Limited Demand for Content Moderation on Social Media”. Only the “Target LGBTQ” is shown in the paper as Figure 1. The other figures’ replication are in the scripts folder in Github, please check the link in the first page.

Different from the original paper, we focused on the 3 social factors groups’ difference of preference on content moderation by treatment. By using the percentages’ differences, we avoids the issue of

4.2 What we acquire

In the paper, the result represents a trend that users prefer to do nothing to the toxic speech even if they may be affected negatively. The first study indicates that people are more likely to react to a toxic speech by reducing how many people can see the post or suspending the account if the target is LGBTQ and there is not much difference if the targets are billionaires or Christians. The second study implies that Democrats may tend to protect LGBTQ more while Republicans care about Christians more. In contrast with these, our reproduction acquires more detailed conclusions from three slightly different perspectives. From gender sight, women show a stronger tendency to protect LGBTQ than men do. From the education level perspective, those who are high school graduates tend to exhibit more protective behaviors when dealing with higher levels of toxic speech while those who are postgraduate act more sensitively to the uncivil levels. From the race perspective, Blacks are usually more sensitive to the LGBTQ target overall.

4.3 What we learn: Point 1: Diverse Consequences of Different Toxic Speech Types

The study challenges past research by revealing various types of toxic speech, including incivility, intolerance, and violent threats. Unlike previous work that often focused solely on labeling manifestations of incivility, this research connects toxic speech types to their consequences. It introduces a nuanced perspective, suggesting that users perceive these speech types as distinct constructs. While intolerance and incivility prompt similar content moderation responses, the study highlights the empirical insight that users view them differently, considering incivility a matter of tone and intolerance a matter of substance, such as discrimination.

4.4 Point 2: Limited Support for Content Moderation

A significant finding is the overall low support for content moderation of uncivil and intolerant content. The majority of respondents express the view that such content should remain online, with censorious forms of moderation, like banning users or removing content, being among the least favored options. Even when presented with extreme cases of toxic speech, such as attacks on the LGBTQ community, a large portion of respondents do not advocate for content moderation. This raises concerns about the broader implications for public discourse, as a substantial portion of users seems reluctant to endorse content moderation, even in the face of highly objectionable speech.

4.5 Point 3: Partisan Consistency and New Research Avenues

Contrary to expectations in an era of affective polarization, the study finds limited evidence that users view moderation of toxic speech through partisan lenses. Democrats, in general, show a greater tendency to demand moderation, but the identity of the victim (Republican or Democrat) does not significantly influence partisans' views. This finding opens up a new research puzzle, suggesting that Americans' strong belief in the value of freedom of speech might be a driving factor. The study calls for further exploration into whether this trend persists in other countries with different legal frameworks. The results emphasize the need to understand content moderation preferences beyond partisan lines and suggest that Americans, despite political polarization, exhibit consistency in their views on this matter.

4.6 Weaknesses

The primary limitation of our research stems from the lack of modeling. The absence of models means that we do not account for potential interactions between different social factors and how they may collectively influence moderation behaviors. This gap is significant because different social groups may have varying degrees of influence on content moderation decisions, and without proper weighting in our analysis, the results might not fully reflect the complexities of these interactions.

Moreover, another limitation is related to the representativeness of our sample. The percentages of different social factor groups in our survey may not accurately mirror the actual composition of society. This discrepancy can lead to skewed results, as the moderation tendencies we observe may be over or under-represented due to sampling bias. For example, if a social factor group that is more likely to advocate for stricter moderation is under-represented in the survey, our findings may underestimate the desire for such moderation across the population.

The survey results involve the interpretation of social tendencies toward content moderation, which are inherently nuanced and subject to the cultural and social dynamics within each group.

However, our current approach does not allow us to explore these dynamics in depth, which could provide valuable insights into the reasons behind the observed moderation preferences.

4.7 Next steps

To address these limitations, our next steps involve the introduction of statistical models, such as regression analyses, which can help us understand the weight and impact of each social factor on moderation decisions. Regression models would allow us to control for various covariates and examine the independent effect of each variable. By doing so, we can also investigate interaction effects, providing a more nuanced understanding of how multiple social factors may interplay to influence content moderation behaviors.

Additionally, efforts should be made to ensure that our sample is more reflective of society's actual demographic composition. This could involve stratified sampling or weighting survey responses to match the demographic makeup of the population. Implementing these methods will likely result in more generalizable and accurate findings that can better inform content moderation policies and practices.

A Appendix

Table 2: Overview of survey questions and variables

Variable	Question	Response
Dependent Variable		
... Support of any form of moderation	In your view, how should social media companies like Facebook and Twitter handle the post above?	Leave it, do nothing (1), Place a warning label on the post (2), Reduce how many people can see the post (3), Permanently remove the post (4), Suspend the person's account (5)
Other Variables		
... Political Identity	Generally speaking, do you consider yourself as being a Republican, a Democrat or an Independent?	Strong Democrat (1), Democrat (2), Leaning Democrat (3), Independent (4), Leaning Republican (5), Republican (6), Strong Republican (7)
... Social media visits	Overall, how often would you say you visit social media platforms (Twitter, Facebook, etc.)?	Every day (1) At least once a week but not every day (2), A few times a month (3), Less often (4)
... Age	What is your age?	[Open text box]
... Gender	What is your gender?	Female (1), Male (2), Other (3)
... Education	What is your highest level of educational qualification?	Less than high school (1), High school graduate (2), Professional degree (3), College (4), Postgraduate (e.g., Masters) (5), PhD (6)
... Race/Ethnicity	What is your ethnicity?	White (1), Black or African American (2), American Indian or Alaska Native (3), Asian (4), Native Hawaiian or Pacific Islander (5), Hispanic (6), Other (7)
... Attention Check	Please indicate your agreement with the following statement below. For our survey, it is essential that participants pay attention. To show us that you are reading this, please select both "Somewhat agree" and "Strongly agree" here.	Strongly disagree (1), Somewhat disagree (2), Neither agree nor disagree (3), Somewhat agree (4), Strongly agree (5) (Multiple selection is possible)
... Emotions	Which of the following emotions best describe your feelings about this social media post?	Anger (Slider 0-100), Enthusiasm (Slider 0-100), Disgust (Slider 0-100), Fear (Slider 0-100), Happiness (Slider 0-100)
... Preference for removing the post	Overall, how strongly do you feel this post should be kept or removed?	Slider 0 (Keep the post) - 100 (Remove the post)

Table 3: Actual Data of Gender Analysis

treatment	gender	handle	percentage
control	Female	Leave it, do nothing	73.66
control	Female	Permanently remove the post	3.76
control	Female	Place a warning label on the post	16.13
control	Female	Reduce how many people can see the post	4.30
control	Female	Suspend the person's account	2.15
control	Male	Leave it, do nothing	79.50
control	Male	Permanently remove the post	2.50
control	Male	Place a warning label on the post	11.00
control	Male	Reduce how many people can see the post	5.50
control	Male	Suspend the person's account	1.50
intolerant	Female	Leave it, do nothing	45.36
intolerant	Female	Permanently remove the post	13.66
intolerant	Female	Place a warning label on the post	29.51
intolerant	Female	Reduce how many people can see the post	5.46
intolerant	Female	Suspend the person's account	6.01
intolerant	Male	Leave it, do nothing	54.00
intolerant	Male	Permanently remove the post	8.50
intolerant	Male	Place a warning label on the post	25.00
intolerant	Male	Reduce how many people can see the post	5.50
intolerant	Male	Suspend the person's account	7.00
non-group-related	Female	Leave it, do nothing	98.45
non-group-related	Female	Place a warning label on the post	1.55
non-group-related	Male	Leave it, do nothing	97.86
non-group-related	Male	Place a warning label on the post	1.60
non-group-related	Male	Reduce how many people can see the post	0.53
threatening	Female	Leave it, do nothing	14.85
threatening	Female	Permanently remove the post	28.71
threatening	Female	Place a warning label on the post	25.25
threatening	Female	Reduce how many people can see the post	3.47
threatening	Female	Suspend the person's account	27.72
threatening	Male	Leave it, do nothing	26.82
threatening	Male	Permanently remove the post	20.11
threatening	Male	Place a warning label on the post	24.02
threatening	Male	Reduce how many people can see the post	4.47
threatening	Male	Suspend the person's account	24.58
uncivil	Female	Leave it, do nothing	47.57
uncivil	Female	Permanently remove the post	14.05
uncivil	Female	Place a warning label on the post	27.57
uncivil	Female	Reduce how many people can see the post	7.03
uncivil	Female	Suspend the person's account	3.78
uncivil	Male	Leave it, do nothing	63.92

uncivil	Male	Permanently remove the post	6.70
uncivil	Male	Place a warning label on the post	20.10
uncivil	Male	Reduce how many people can see the post	4.64
uncivil	Male	Suspend the person’s account	4.64

Table 4: Actual Data of Race Analysis

treatment	race	handle	percentage
control	Black	Leave it, do nothing	65.52
control	Black	Permanently remove the post	3.45
control	Black	Place a warning label on the post	27.59
control	Black	Reduce how many people can see the post	3.45
control	Hispanic	Leave it, do nothing	61.29
control	Hispanic	Permanently remove the post	9.68
control	Hispanic	Place a warning label on the post	19.35
control	Hispanic	Reduce how many people can see the post	6.45
control	Hispanic	Suspend the person’s account	3.23
control	Others	Leave it, do nothing	73.68
control	Others	Permanently remove the post	1.75
control	Others	Place a warning label on the post	12.28
control	Others	Reduce how many people can see the post	7.02
control	Others	Suspend the person’s account	5.26
control	White	Leave it, do nothing	80.37
control	White	Permanently remove the post	2.59
control	White	Place a warning label on the post	11.48
control	White	Reduce how many people can see the post	4.44
control	White	Suspend the person’s account	1.11
intolerant	Black	Leave it, do nothing	33.33
intolerant	Black	Permanently remove the post	11.11
intolerant	Black	Place a warning label on the post	36.11
intolerant	Black	Reduce how many people can see the post	11.11
intolerant	Black	Suspend the person’s account	8.33
intolerant	Hispanic	Leave it, do nothing	56.67
intolerant	Hispanic	Permanently remove the post	3.33
intolerant	Hispanic	Place a warning label on the post	20.00
intolerant	Hispanic	Reduce how many people can see the post	3.33
intolerant	Hispanic	Suspend the person’s account	16.67
intolerant	Others	Leave it, do nothing	52.27
intolerant	Others	Permanently remove the post	6.82
intolerant	Others	Place a warning label on the post	31.82
intolerant	Others	Reduce how many people can see the post	4.55
intolerant	Others	Suspend the person’s account	4.55
intolerant	White	Leave it, do nothing	50.18

intolerant	White	Permanently remove the post	12.73
intolerant	White	Place a warning label on the post	25.82
intolerant	White	Reduce how many people can see the post	5.45
intolerant	White	Suspend the person’s account	5.82
non-group-related	Black	Leave it, do nothing	93.75
non-group-related	Black	Place a warning label on the post	6.25
non-group-related	Hispanic	Leave it, do nothing	96.97
non-group-related	Hispanic	Reduce how many people can see the post	3.03
non-group-related	Others	Leave it, do nothing	97.96
non-group-related	Others	Place a warning label on the post	2.04
non-group-related	White	Leave it, do nothing	98.90
non-group-related	White	Place a warning label on the post	1.10
threatening	Black	Leave it, do nothing	10.53
threatening	Black	Permanently remove the post	26.32
threatening	Black	Place a warning label on the post	26.32
threatening	Black	Suspend the person’s account	36.84
threatening	Hispanic	Leave it, do nothing	10.00
threatening	Hispanic	Permanently remove the post	40.00
threatening	Hispanic	Place a warning label on the post	20.00
threatening	Hispanic	Suspend the person’s account	30.00
threatening	Others	Leave it, do nothing	22.22
threatening	Others	Permanently remove the post	22.22
threatening	Others	Place a warning label on the post	31.48
threatening	Others	Reduce how many people can see the post	7.41
threatening	Others	Suspend the person’s account	16.67
threatening	White	Leave it, do nothing	21.16
threatening	White	Permanently remove the post	23.89
threatening	White	Place a warning label on the post	23.21
threatening	White	Reduce how many people can see the post	3.75
threatening	White	Suspend the person’s account	27.99
uncivil	Black	Leave it, do nothing	40.74
uncivil	Black	Permanently remove the post	18.52
uncivil	Black	Place a warning label on the post	29.63
uncivil	Black	Reduce how many people can see the post	3.70
uncivil	Black	Suspend the person’s account	7.41
uncivil	Hispanic	Leave it, do nothing	40.00
uncivil	Hispanic	Permanently remove the post	24.00
uncivil	Hispanic	Place a warning label on the post	16.00
uncivil	Hispanic	Reduce how many people can see the post	8.00
uncivil	Hispanic	Suspend the person’s account	12.00
uncivil	Others	Leave it, do nothing	47.37
uncivil	Others	Permanently remove the post	7.02
uncivil	Others	Place a warning label on the post	29.82
uncivil	Others	Reduce how many people can see the post	10.53

uncivil	Others	Suspend the person's account	5.26
uncivil	White	Leave it, do nothing	60.51
uncivil	White	Permanently remove the post	9.42
uncivil	White	Place a warning label on the post	22.10
uncivil	White	Reduce how many people can see the post	5.07
uncivil	White	Suspend the person's account	2.90

Table 5: Actual Data of Education Level Analysis

treatment	educ	handle	percentage
control	College	Leave it, do nothing	75.96
control	College	Permanently remove the post	2.88
control	College	Place a warning label on the post	13.46
control	College	Reduce how many people can see the post	5.29
control	College	Suspend the person's account	2.40
control	High school graduate	Leave it, do nothing	81.13
control	High school graduate	Permanently remove the post	2.83
control	High school graduate	Place a warning label on the post	11.32
control	High school graduate	Reduce how many people can see the post	3.77
control	High school graduate	Suspend the person's account	0.94
control	Postgraduate	Leave it, do nothing	77.78
control	Postgraduate	Permanently remove the post	2.22
control	Postgraduate	Place a warning label on the post	11.11
control	Postgraduate	Reduce how many people can see the post	8.89
intolerant	College	Leave it, do nothing	47.15
intolerant	College	Permanently remove the post	10.88
intolerant	College	Place a warning label on the post	29.02
intolerant	College	Reduce how many people can see the post	7.25
intolerant	College	Suspend the person's account	5.70
intolerant	High school graduate	Leave it, do nothing	52.34
intolerant	High school graduate	Permanently remove the post	7.48
intolerant	High school graduate	Place a warning label on the post	27.10
intolerant	High school graduate	Reduce how many people can see the post	5.61
intolerant	High school graduate	Suspend the person's account	7.48
intolerant	Postgraduate	Leave it, do nothing	54.10
intolerant	Postgraduate	Permanently remove the post	14.75
intolerant	Postgraduate	Place a warning label on the post	22.95
intolerant	Postgraduate	Reduce how many people can see the post	1.64
intolerant	Postgraduate	Suspend the person's account	6.56
non-group-related	College	Leave it, do nothing	98.33
non-group-related	College	Place a warning label on the post	1.11
non-group-related	College	Reduce how many people can see the post	0.56
non-group-related	High school graduate	Leave it, do nothing	98.11

non-group-related	High school graduate	Place a warning label on the post	1.89
non-group-related	Postgraduate	Leave it, do nothing	100.00
threatening	College	Leave it, do nothing	20.11
threatening	College	Permanently remove the post	24.87
threatening	College	Place a warning label on the post	23.28
threatening	College	Reduce how many people can see the post	3.70
threatening	College	Suspend the person’s account	28.04
threatening	High school graduate	Leave it, do nothing	21.43
threatening	High school graduate	Permanently remove the post	19.64
threatening	High school graduate	Place a warning label on the post	21.43
threatening	High school graduate	Reduce how many people can see the post	5.36
threatening	High school graduate	Suspend the person’s account	32.14
threatening	Postgraduate	Leave it, do nothing	12.00
threatening	Postgraduate	Permanently remove the post	34.00
threatening	Postgraduate	Place a warning label on the post	32.00
threatening	Postgraduate	Reduce how many people can see the post	2.00
threatening	Postgraduate	Suspend the person’s account	20.00
uncivil	College	Leave it, do nothing	55.38
uncivil	College	Permanently remove the post	7.53
uncivil	College	Place a warning label on the post	24.19
uncivil	College	Reduce how many people can see the post	8.06
uncivil	College	Suspend the person’s account	4.84
uncivil	High school graduate	Leave it, do nothing	52.63
uncivil	High school graduate	Permanently remove the post	13.16
uncivil	High school graduate	Place a warning label on the post	27.19
uncivil	High school graduate	Reduce how many people can see the post	4.39
uncivil	High school graduate	Suspend the person’s account	2.63
uncivil	Postgraduate	Leave it, do nothing	64.91
uncivil	Postgraduate	Permanently remove the post	12.28
uncivil	Postgraduate	Place a warning label on the post	14.04
uncivil	Postgraduate	Reduce how many people can see the post	5.26
uncivil	Postgraduate	Suspend the person’s account	3.51

References

Mills, Blake Robert. 2022. *MetBrewer: Color Palettes Inspired by Works at the Metropolitan Museum of Art*. https://CRAN.R-project.org/package=MetBrewer.

PRADEL F, KOSMIDIS S, ZILINSKY J. n.d. "Toxic Speech and Limited Demand for Content Moderation on Social Media." CAmerican Political Science Review. https://doi.org/10.1017/S000305542300134X.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

———. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Vogels, Emily A. 2021. *The State of Online Harassment*. Pew Rsearch Center. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2021/01/PI_2021.01.13_Online-Harassment_FINAL-1.pdf.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2023. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.