Supplementary Information for

Toxic Speech and Limited Demand for Content Moderation on Social Media

This PDF file includes:

Tables S1 to S22
Figures S1 to S4
Content moderation practices and public opinion in the US and beyond
Further analysis of incivility and intolerance as distinct constructs
Further analysis of heterogeneous effects

List of Tables

List of Figures

Study I - Social groups

Table S1: Overview of manipulated social media posts in Study I (social groups)

	a) Study - Target LGBTQ
Control A: No group mentioned	i just watched a movie everyone is talking about – can't say it was great..
Control B: Anti-target	maybe i am old school but i think hollywood should stick to making movies and stop pushing stories about gay couples
T1: uncivil treatment	hollywood should stick to making movies! i don't need to be bombarded by gay politics! just fuck off!
T2: intolerant treatment	hollywood should stick to making movies. i don't need to be bombarded by gay politics. no homos on TV or next to me or anywhere!
T3: threatening treatment	hollywood should stick to making movies! i don't need to be bombarded by gay politics. next time i see one of you fags i'll smash your face!***
	b) Study - Target Billionaires
Control A: No group mentioned	i just watched a movie everyone is talking about – can't say it was great..
Control B: Anti-target	enough with the non-stop coverage of billionaires and their lifestyles. . .
T1: uncivil treatment	enough with the non-stop coverage of billionaires and their lifestyles . . . just fuck off!
T2: intolerant treatment	enough with the non-stop coverage of billionaires and their lifestyles . . . they should keep quiet. let's abolish them!
T3: threatening treatment	enough with the non-stop coverage of billionaires and their lifestyle ...if i ever see one, I'LL PUNCH THEM IN THE FACE!
	c) Study - Target Christians
Control A: No group mentioned	I just watched the movie everyone is talking about – can't say it was great..
Control B: Anti-target	Was just driving behind this guy with a collection of JESUS SAVES! bumper stickers on his pickup truck.. 👎👎
T1: uncivil treatment	Was just driving behind this guy with a collection of JESUS SAVES! bumper stickers on his pickup truck.. Tell you what. FUCK THESE RELIGIOUS FANATICS!!!
T2: intolerant treatment	Was just driving behind this guy with a collection of JESUS SAVES! bumper stickers on his pickup truck.. Tell you what. Why don't you take your pickup truck and drive ELSEWHERE? Your kind is not welcome here.
T3: threatening treatment	Was just driving behind this guy with a collection of JESUS SAVES! bumper stickers on his pickup truck.. Tell you what. next time i see one of them I will PUNCH THEM IN THE FACE

****Note:* Because in this treatment the threat is phrased differently compared to the other experiments ("I'll smash your face" is used instead of "I will punch them in the face"), in our replication study done after Musk's takeover of Twitter (please see SI, page 10), we extended this experiment (target: LGBTQ) by also exposing users to a treatment with the exact same wording as in all the other experiments. For the relevant analysis and results please see "threatening post(new)" in Figure S3, page 11 in the SI.

Table S2: Overview of survey questions and variables

Variable	Question	Response categories
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 Key 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)
... Perception of the post (best description of the post)	Considering the post above, which of the following features do you think fits best as a description of the social media post? Please order the features, with 1 being the best fit. To rank the listed items, drag and drop each item.	uncivil (1), intolerant (2), threatening (3), civil (4) (randomized order)
... 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)

Table S3: Sociodemographics of participants - LGBTQ study

Variable	N	Percent
Gender	1936	
... Female	950	49.1%
... Male	960	49.6%
... Other	26	1.3%
Age Group	1934	
... Born 1991 or later	737	38.1%
... Born 1975-1990	715	37%
... Born 1959-1974	343	17.7%
... Born 1943-1958	138	7.1%
... Born 1927-1942	1	0.1%
... Born 1911-1926	0	0%
Race and Ethnicity	1930	
... Black	143	7.4%
... Hispanic	139	7.2%
... Race other/multiple	261	13.5%
... White	1387	71.9%
Political Identity	1936	
... Democrat	1021	52.7%
... Independent	534	27.6%
... Republican	381	19.7%
Education	1936	
... College	956	49.4%
... High school graduate	545	28.2%
... Less than high school	16	0.8%
... PhD	41	2.1%
... Postgraduate (e.g. Masters)	271	14%
... Professional degree	88	4.5%

Table S4: Sociodemographics of participants - Billionaires study

Variable	N	Percent
Gender	1860	
... Female	912	49%
... Male	910	48.9%
... Other	38	2%
Age Group	1856	
... Born 1991 or later	839	45.2%
... Born 1975-1990	649	35%
... Born 1959-1974	286	15.4%
... Born 1943-1958	77	4.1%
... Born 1927-1942	5	0.3%
... Born 1911-1926	0	0%
Race and Ethnicity	1854	
... Black	158	8.5%
... Hispanic	172	9.3%
... Race other/multiple	262	14.1%
... White	1262	68.1%
Political Identity	1860	
... Democrat	1033	55.5%
... Independent	485	26.1%
... Republican	342	18.4%
Education	1860	
... College	892	48%
... High school graduate	536	28.8%
... Less than high school	10	0.5%
... PhD	35	1.9%
... Postgraduate (e.g. Masters)	280	15.1%
... Professional degree	79	4.2%

Table S5: Sociodemographics of participants - Christian study

Variable	N	Percent
Gender	1334	
... Female	658	49.3%
... Male	652	48.9%
... Other	24	1.8%
Age Group	1331	
... Born 1991 or later	661	49.7%
... Born 1975-1990	501	37.6%
... Born 1959-1974	146	11%
... Born 1943-1958	21	1.6%
... Born 1927-1942	2	0.2%
... Born 1911-1926	0	0%
Race and Ethnicity	1329	
... Black	120	9%
... Hispanic	120	9%
... Race other/multiple	162	12.2%
... White	927	69.8%
Political Identity	1334	
... Democrat	792	59.4%
... Independent	342	25.6%
... Republican	200	15%
Education	1334	
... College	662	49.6%
... High school graduate	424	31.8%
... Less than high school	17	1.3%
... PhD	14	1%
... Postgraduate (e.g. Masters)	151	11.3%
... Professional degree	48	3.6%

Table S6: Sociodemographic characteristics of participants of the pooled data in comparison to sociodemographic characteristics of participants in the 2020 wave of the American National Election Studies (ANES)

Variable	N (Pooled Data)	Percent (Pooled Data)	N (ANES 2020)	Percent (ANES 2020)	Dif. Percent (Pooled-ANES)
Gender	5130		8226		
... Female	2520	49.1%	4262	52%	-2.9%
... Male	2522	49.2%	3964	48%	1.2%
Age Group	5121		7951		
... Born 1991 or later	2237	43.7%	1484	19%	24.7%
... Born 1975-1990	1865	36.4%	2076	26%	10.4%
... Born 1959-1974	775	15.1%	2186	27%	-11.9%
... Born 1943-1958	236	4.6%	1792	23%	-18.4%
... Born 1927-1942	8	0.2%	404	5%	-4.8%
... Born 1911-1926	0	0%	8	0%	0%
Race and Ethnicity	5113		8198		
... Black	421	8.2%	935	11%	-2.8%
... Hispanic	431	8.4%	1108	14%	-5.6%
... Race other/multiple	685	13.4%	773	9%	4.4%
... White	3576	69.9%	5383	66%	3.9%
Political Identity	5130		8251		
... Democrat	2846	55.5%	3808	46%	9.5%
... Independent	1361	26.5%	976	12%	14.5%
... Republican	923	18%	3467	42%	-24%
Education (Pooled data)	5130				
... College	2510	48.9%			
... High school graduate	1505	29.3%			
... Less than high school	43	0.8%			
... PhD	90	1.8%			
... Postgraduate (e.g. Masters)	702	13.7%			
... Professional degree	215	4.2%			
Education (ANES)			8147		
... Less than high school graduate			98	1%	
... High School (Grades 9-12)			2687	33%	
... Some College, no Degree			2376	29%	
... College Degree/ Post-grad			2986	37%	

Table S7: Percentage of participants preferring any form of moderation by treatment groups - Study I (social groups)

group	LGBTQ (%)	Billionaires (%)	Christians (%)	Pooled data (%)
all	40	14	24	26
Control A: No group mentioned	2	3	1	2
Control B: Anti-target	23	3	9	12
T1: uncivil post	44	19	31	32
T2: intolerant	51	12	31	31
T3: threatening	80	34	45	54
Observations (N)	1936	1860	1334	5130

Note: We considered that participants prefer any form of moderation if they selected any of “Permanently remove the post”, “Place a warning label on the post”, “Reduce how many people can see the post”, or “Suspend the person’s account” as their preferred action against the shown post.

Table S8: Preferences for type of moderation by treatment groups for all experiments in Study I (social groups)

treatment	how to handle the post	Pooled (%)	LGBTQ (%)	Billionaires (%)	Christians (%)
Control A: No group mentioned	Leave it, do nothing	98.0	98.2	97.3	98.5
	Place a warning label on the post	1.2	1.6	1.1	0.8
	Reduce how many people can see the post	0.8	0.3	1.4	0.8
	Suspend the person’s account	0.1		0.3	
Control B: Anti-target	Leave it, do nothing	87.7	76.9	96.8	90.8
	Permanently remove the post	1.6	3.1	0.3	1.5
	Place a warning label on the post	6.9	13.4	1.1	5.5
	Reduce how many people can see the post	3.1	4.9	1.9	2.2
	Suspend the person’s account	0.7	1.8		
T1: uncivil post	Leave it, do nothing	68.3	56.0	81.1	68.5
	Permanently remove the post	6.6	10.6	3.3	5.2
	Place a warning label on the post	18.6	23.3	11.8	21.0
	Reduce how many people can see the post	4.3	6.0	3.3	3.4
	Suspend the person’s account	2.3	4.2	0.6	1.9
T2: intolerant post	Leave it, do nothing	68.7	49.4	88.0	69.0
	Permanently remove the post	6.7	11.1	1.8	7.5
	Place a warning label on the post	16.6	27.1	7.3	14.5
	Reduce how many people can see the post	4.9	5.7	2.6	7.1
	Suspend the person’s account	3.1	6.7	0.3	2.0
T3: threatening post	Leave it, do nothing	46.0	20.2	66.4	54.8
	Permanently remove the post	13.3	24.6	5.7	7.7
	Place a warning label on the post	22.3	24.4	18.3	25.0
	Reduce how many people can see the post	5.4	3.9	6.5	6.2
	Suspend the person’s account	12.9	26.9	3.2	6.2
Observations (N)		5130	1936	1860	1334

Table S9: Logit models underlying the marginal effects displayed in Figure 2. Cell entries are regression coefficients (not exponentiated) and standard errors are shown in parentheses.

	Dependent variable:			
	Any Moderation			
	Pooled results	LGBTQ target	Billionaire target	Christian target
Uncivil post	1.322	0.961	1.942	1.512
	(0.122)	(0.158)	(0.322)	(0.248)
Intolerant post	1.317	1.226	1.410	1.489
	(0.122)	(0.157)	(0.333)	(0.250)
Threatening post	2.414	2.574	2.717	2.099
	(0.122)	(0.175)	(0.313)	(0.243)
No group mentioned (movie only)	-1.977	-2.796	-0.185	-1.899
	(0.243)	(0.400)	(0.435)	(0.546)
Anti-target post	(ref. group)	(ref. group)	(ref. group)	(ref. group)
Respondent PID: Democrat	0.587			
	(0.089)			
Respondent PID: Republican	-0.245			
	(0.119)			
Respondent PID: Independent	(ref. group)			
Study: Christians	(ref. group)			
Study: Billionaires	-0.744			
	(0.101)			
Study: LGBTQ	1.034			
	(0.091)			
(Intercept)	-2.618	-1.201	-3.398	-2.291
	(0.137)	(0.120)	(0.293)	(0.210)
Num.Obs.	5,130	1,936	1,860	1,334
Log Likelihood	-2222.501	-972.741	-653.869	-615.669
Akaike Inf. Crit.	4463.0	1955.5	1317.7	1241.3

Note: The dependent variable “Any Moderation” takes the value of 1 if participants indicate support for any form of moderation, and 0 if participants responded “Leave it, do nothing”.

Table S10: Logit models underlying the contrasts displayed in Figure 4. Cell entries are coefficients (not exponentiated, left column) and standard errors (right column).

	Dependent variable:	
	Any Moderation	
	Coef. estimate	Std. error
No group mentioned (movie only)	-12.986	180.520
Uncivil	1.175	0.558
Intolerant	1.258	0.566
Threatening	1.959	0.518
Republican	0.470	0.704
Democrat	0.356	0.533
Billionaires	-0.832	0.748
LGBTQ	0.900	0.527
No group x Republican	12.323	180.523
Uncivil x Republican	-0.541	0.863
Intolerant x Republican	-0.083	0.853
Threatening x Republican	-0.325	0.799
No group x Democrat	10.786	180.522
Uncivil x Democrat	0.639	0.638
Intolerant x Democrat	0.357	0.646
Threatening x Democrat	0.407	0.605
No group x Billionaires	13.112	180.522
Uncivil x Billionaires	0.396	0.863
Intolerant x Billionaires	0.041	0.876
Threatening x Billionaires	0.371	0.822
No group x LGBTQ	10.725	180.522
Uncivil x LGBTQ	0.125	0.644
Intolerant x LGBTQ	0.366	0.649
Threatening x LGBTQ	0.930	0.622
Republican x Billionaires	-0.599	1.164
Democrat x Billionaires	-0.291	0.881
Republican x LGBTQ	-0.938	0.847
Democrat x LGBTQ	0.564	0.608
No group x Republican x Billionaires	-12.293	180.527
Uncivil x Republican x Billionaires	1.369	1.334
Intolerant x Republican x Billionaires	0.302	1.347
Threatening x Republican x Billionaires	0.927	1.274
No group x Democrat x Billionaires	-11.446	180.525
Uncivil x Democrat x Billionaires	-0.267	1.009
Intolerant x Democrat x Billionaires	-0.227	1.027
Threatening x Democrat x Billionaires	0.046	0.970
No group x Republican x LGBTQ	-10.804	180.525
Uncivil x Republican x LGBTQ	0.420	1.035
Intolerant x Republican x LGBTQ	-0.829	1.034
Threatening x Republican x LGBTQ	-0.215	0.984
No group x Democrat x LGBTQ	-13.091	180.526
Uncivil x Democrat x LGBTQ	-1.090	0.743
Intolerant x Democrat x LGBTQ	-0.651	0.750
Threatening x Democrat x LGBTQ	-0.248	0.752
(Intercept)	-2.580	0.464
Num. Obs.	5130	
Log Likelihood	-2156.855	
Akaike Inf. Crit.	4403.7	

Note: Reference categories = Type of post: Anti-target post, Political Identity = Independent Respondent, Study = Study: Christians. The dependent variable "Any Moderation" takes the value of 1 if participants indicate support for any form of moderation, and 0 if participants responded "Leave it, do nothing".

Table S11: Estimates and 95% confidence intervals displayed in Figure 4 (top and bottom panel) - Study I (social groups)

		Dependent variable:		
		Any Moderation		
Treatment group	Target	$\Delta_{Pr(Dem-Rep)}$	p-value	CI
Uncivil post	LGBTQ	0.247	0.000	[0.128,0.366]
Anti-target but without hostility	LGBTQ	0.214	0.000	[0.116,0.311]
Threatening post	LGBTQ	0.358	0.000	[0.242,0.474]
Intolerant post	LGBTQ	0.446	0.000	[0.337,0.556]
Intolerant post	Billionaires	0.011	0.797	[-0.074,0.096]
Uncivil post	Billionaires	-0.045	0.462	[-0.164,0.075]
Threatening post	Billionaires	0.010	0.875	[-0.118,0.138]
Anti-target but without hostility	Billionaires	0.006	0.803	[-0.04,0.052]
Threatening post	Christians	0.152	0.064	[-0.009,0.314]
Anti-target but without hostility	Christians	-0.011	0.851	[-0.12,0.099]
Intolerant post	Christians	0.070	0.389	[-0.089,0.23]
Uncivil post	Christians	0.213	0.003	[0.074,0.352]

			Dependent variable:	
			Any Moderation	
Treatment group	Target	Respondents' PID	Estimate	CI
Effect of intolerant language	Christians	Republican	0.174	[0.001,0.347]
Effect of intolerant language	Billionaires	Republican	0.089	[0.007,0.170]
Effect of intolerant language	LGBTQ	Republican	0.088	[-0.026,0.202]
Effect of intolerant language	Christians	Democrat	0.255	[0.168,0.342]
Effect of intolerant language	Billionaires	Democrat	0.094	[0.042,0.146]
Effect of intolerant language	LGBTQ	Democrat	0.320	[0.228,0.413]
Effect of threatening language	Christians	Republican	0.275	[0.104,0.446]
Effect of threatening language	Billionaires	Republican	0.324	[0.206,0.442]
Effect of threatening language	LGBTQ	Republican	0.446	[0.314,0.577]
Effect of threatening language	Christians	Democrat	0.438	[0.344,0.531]
Effect of threatening language	Billionaires	Democrat	0.328	[0.26,0.397]
Effect of threatening language	LGBTQ	Democrat	0.589	[0.514,0.665]
Effect of uncivil language	Christians	Republican	0.078	[-0.075,0.231]
Effect of uncivil language	Billionaires	Republican	0.214	[0.100,0.327]
Effect of uncivil language	LGBTQ	Republican	0.171	[0.048,0.293]
Effect of uncivil language	Christians	Democrat	0.301	[0.212,0.390]
Effect of uncivil language	Billionaires	Democrat	0.163	[0.104,0.223]
Effect of uncivil language	LGBTQ	Democrat	0.204	[0.110,0.297]
Num. Obs.	5130			
Log Likelihood	-2156.855			
Akaike Inf. Crit.	4403.7			

Note: The top table corresponds to the estimates visualized in the top panel of Figure 4. Each estimate corresponds to partisan differences in the predicted probability to demand moderation with higher values denoting higher demands from Democratic respondents. The estimates appearing in the lower panel of Figure 4 can be found in the lower part of the table and they correspond to partisan differences when treatments are compared to the Anti-target group. The dependent variable "Any Moderation" takes the value of 1 if participants indicate support for any form of moderation, and 0 if participants responded "Leave it, do nothing".

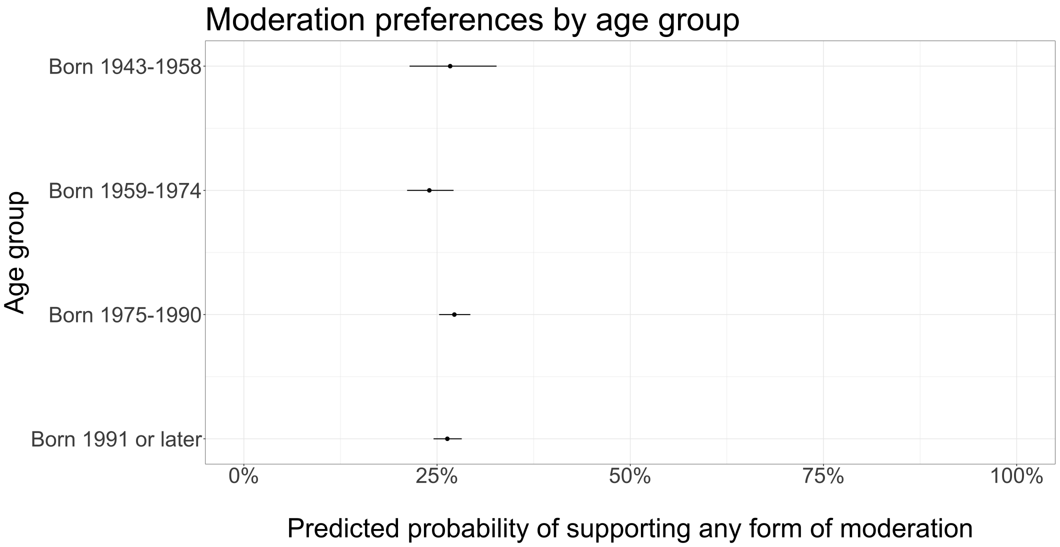


Figure S1: Predicted values and 95% confidence intervals based on binomial logit models predicting participants' support of any form of moderation versus no moderation with participants' age.

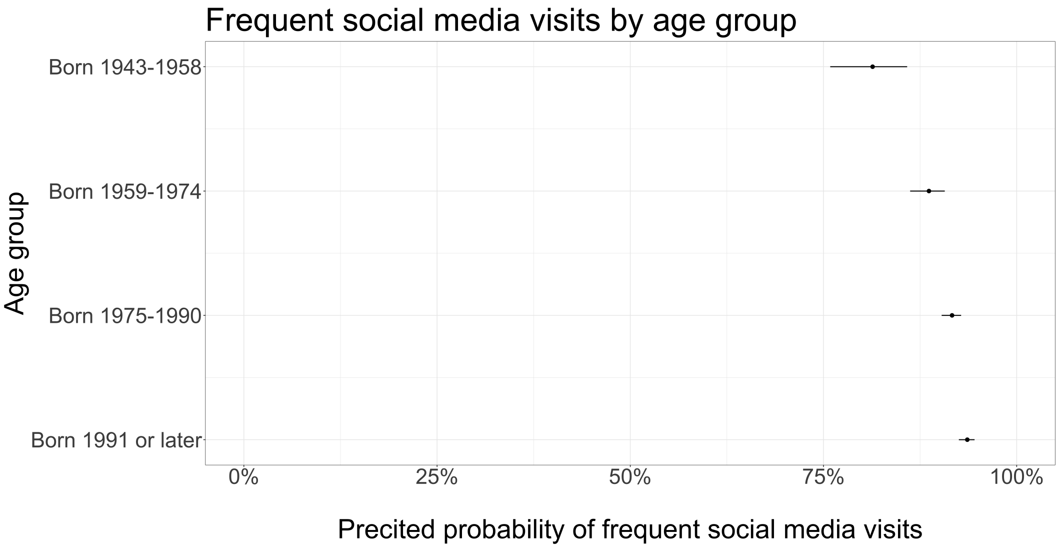


Figure S2: Predicted values and 95% confidence intervals based on binomial logit models predicting participants' self-reported frequency of social media visits with participants' age.

Replication of LGBTQ study after Elon Musk's Twitter take-over

We replicated and expanded upon the previous LGBTQ study. The replication study investigates the effect of toxic speech toward LGBTQ on content moderation preferences. The only difference is the new context brought about by Elon Musk's acquisition of Twitter.

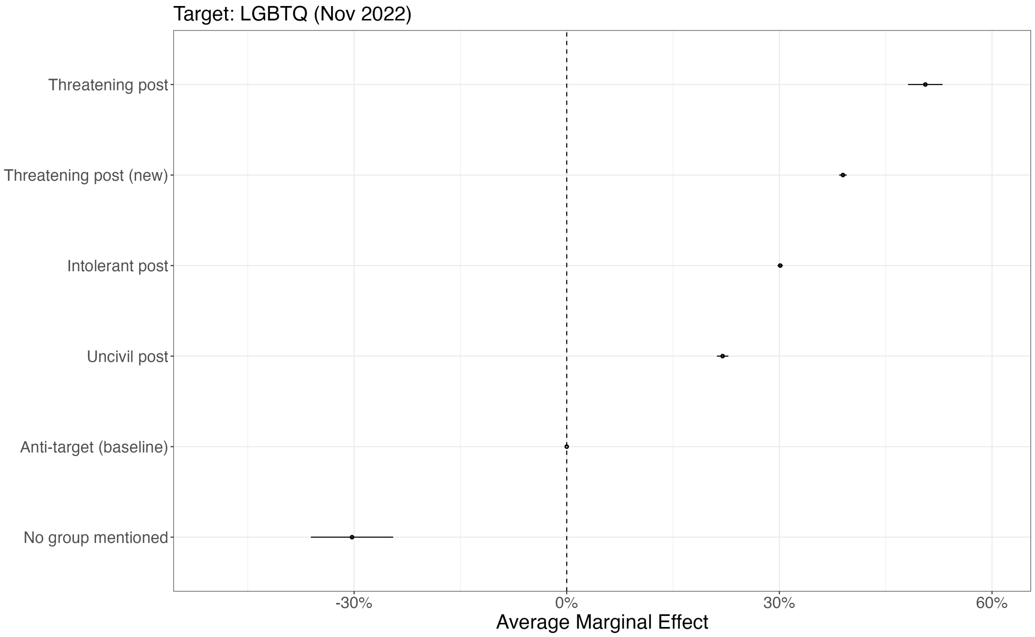
Participants were randomly assigned to one of six conditions (1x6-design). Five of the treatments were exactly the same as in the main LGBTQ study (see also overview of manipulated social media posts with exact wording in Table S1), and one additional treatment was added to make treatments across experiments more comparable. Specifically:

Participants were randomly assigned to one of six conditions in a between-subjects design, i.e., an incivility-factor: in which incivility salience was triggered with a reading task about an uncivil social media post about LGBTQ people (treatment 1: "hollywood should stick to making movies! i don't need to be bombarded by gay couples and their politics! just fuck off!") vs. a reading task about an intolerant social media post about LGBTQ people (treatment 2: "hollywood should stick to making movies. i don't need to be bombarded by gay politics. no homos on TV or next to me or anywhere!") vs. reading task about a threatening social media post about LGBTQ people (treatment 3: "hollywood should stick to making movies! i don't need to be bombarded by gay politics. next time i see one of you fags i'll smash your face!") vs. a similar reading task about a civil social media post about LGBTQ (control 1: "maybe i am old school but i think hollywood should stick to making movies and stop pushing stories about gay couples") vs. a similar reading task about a non-LGBTQ-related civil social media post (control 2: "I just watched a movie everyone is talking about – can't say it was great .."). As an extension of our previous experiment, our sixth treatment used an alternative version of the threat treatment ("hollywood should stick to making movies! i don't need to be bombarded by gay politics.. Tell you what. next time i see one of them I will PUNCH THEM IN THE FACE") allowing us to keep the content across treatments as similar as possible (for more details, see also our pre-registration). As shown below, we also see limited demands for content moderation for our new threat-treatment group. Thus, the implications do not change.

The study was pre-registered on AsPredicted before data collection. The anonymized preregistrations of the studies can be found under the following `link`. We replicated key findings of the manuscript that are shown in Figure 2 and the Figure 3 in the main manuscript. The replicated Figure 2 is shown below as Figure S3 (and Table S12) and the replicated Figure 3 as Figure S4 (and Table S13).

Our results (reported below) also hold when we replicate the experiment two weeks after Elon Musk's takeover of Twitter - a move that created a media spectacle around one of the most central social media platforms for political interactions. While this event initiated a vibrant debate leading many to ask how users would react to a new status quo characterized by lighter content moderation, our findings remain effectively unchanged.

Figure S3: Effects of LGBTQ treatments on support for some form of content moderation in the Post-Musk replication of the LGBTQ study.



Note: The dependent variable is set to 1 if the respondent selected any of "Permanently remove the post", "Place a warning label on the post", "Reduce how many people can see the post", or "Suspend the person's account" as their preferred action against the offending post (Figure 2 in the main manuscript shows original effects). The logit models underlying the marginal effects displayed in Figure 2 and Figure S3 are shown in Table S12 in the SI.

Table S12: Logit models predicting support for any form of content moderation and 95% confidence intervals - Post-Musk replication of LGBTQ study. Logit models underlying the marginal effects displayed in Figure 2 and Figure S3.

	Dependent variable:	
	Any Moderation	
	LGBTQ study (July 2022)	LGBTQ study (November 2022) (Replication)
No group mentioned	−2.796*** [−3.580, −2.013]	−2.932*** [−3.868, −1.996]
Uncivil post	0.961*** [0.651, 1.271]	0.909*** [0.501, 1.318]
Intolerant post	1.226*** [0.918, 1.535]	1.246*** [0.832, 1.660]
Threatening post	2.574*** [2.232, 2.916]	2.331*** [1.854, 2.808]
Threatening post (new)		1.650*** [1.220, 2.081]
Constant	−1.201*** [−1.436, −0.965]	−0.716*** [−1.013, −0.419]
Observations	1,936	1,183
Log Likelihood	−972.741	−619.530
Akaike Inf. Crit.	1,955.482	1,251.059

Note: *p<0.05; **p<0.01; ***p<0.001, Reference category = Type of post: Anti-target The dependent variable "Any Moderation" takes the value of 1 if participants indicate support for any form of moderation, and 0 if participants responded "Leave it, do nothing".

Figure S4: Comparison of user preferences in the original LGBTQ study and its replication post-Musk, by treatment and by experiment (Figure 3 in the main manuscript shows original effects of the LGBTQ study compared to the other target studies)

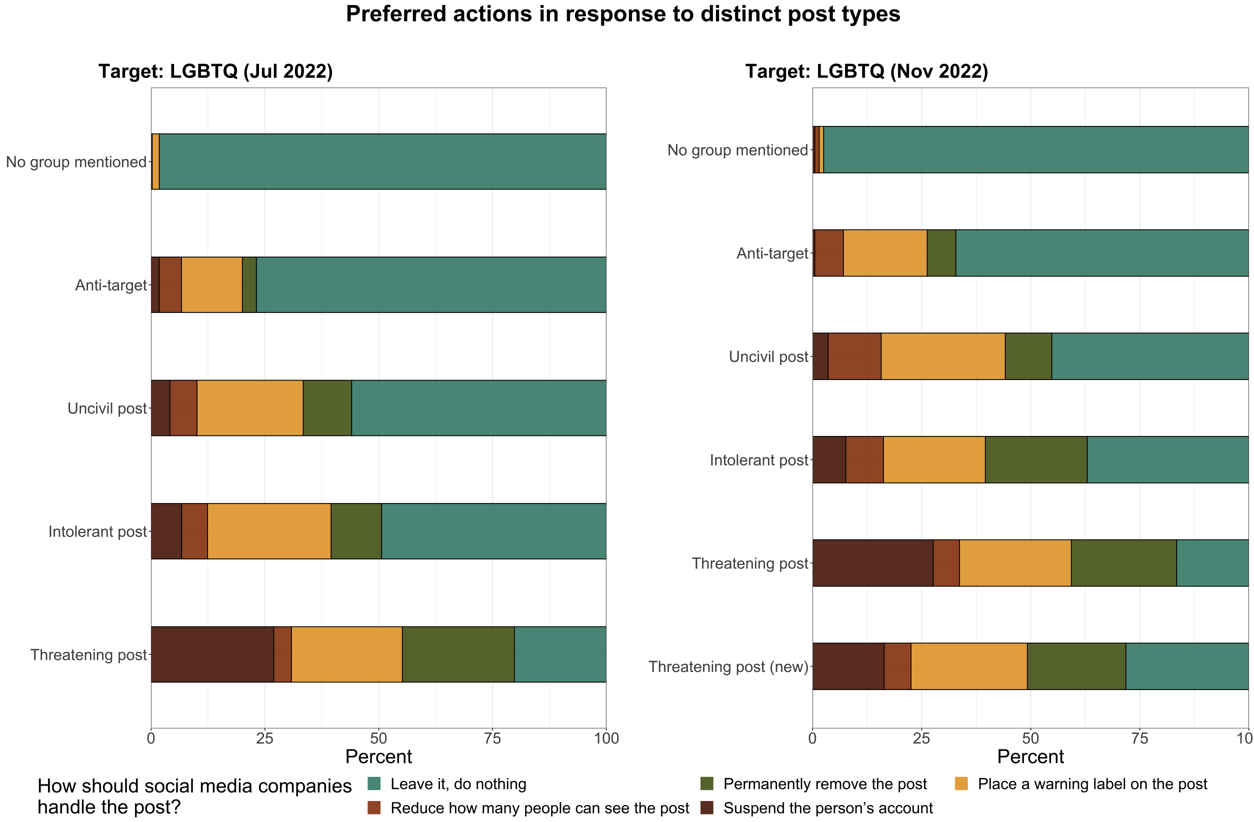


Table S13: Preferences for type of moderation by treatment groups for the original LGBTQ study and the Post-Musk replication of the LGBTQ study

treatment	how to handle the post	LGBTQ (%)	LGBTQ (Replication) (%)
Control A: No group mentioned	Leave it, do nothing	98.2	97.5
	Place a warning label on the post	1.6	1.0
	Reduce how many people can see the post	0.3	1.0
Control B: Anti-target	Leave it, do nothing	76.9	67.2
	Permanently remove the post	3.1	6.6
	Place a warning label on the post	13.4	19.2
	Reduce how many people can see the post	4.9	6.6
	Suspend the person's account	1.8	0.5
T1: uncivil post	Leave it, do nothing	56.0	45.2
	Permanently remove the post	10.6	10.7
	Place a warning label on the post	23.3	28.4
	Reduce how many people can see the post	6.0	12.2
	Suspend the person's account	4.2	3.5
T2: intolerant post	Leave it, do nothing	49.4	37.1
	Permanently remove the post	11.1	23.4
	Place a warning label on the post	27.1	23.4
	Reduce how many people can see the post	5.7	8.6
	Suspend the person's account	6.7	7.6
T3: threatening post	Leave it, do nothing	20.2	16.6
	Permanently remove the post	24.6	24.1
	Place a warning label on the post	24.4	25.6
	Reduce how many people can see the post	3.9	6.0
	Suspend the person's account	26.9	27.6
Observations (N)		1936	1183

Study II - Partisan Identities

Table S14: Overview of manipulated social media posts in Study II (partisans)

	Study: Partisan-Attacks
Control A: No group mentioned	i just watched the movie everyone is talking about – can't say it was great..
Control B: Anti-target	Was just driving behind this guy with a Vote Democrat [Republican] bumper sticker on his car.. 👎👎
T1: uncivil treatment	Was just driving behind this guy with a Vote Democrat [Republican] bumper sticker on his car.. Tell you what. FUCK DEMOCRATS [REPUBLICANS]!
T2: intolerant treatment	Was just driving behind this guy with a Vote Democrat [Republican] bumper sticker on his car.. Tell you what. Why don't you take your car and DRIVE ELSEWHERE? Your kind is not welcome here.
T3: threatening treatment	Was just driving behind this guy with a Vote Democrat [Republican] bumper sticker on his car.. Tell you what. next time I see one of them I WILL PUNCH THEM IN THE FACE

Table S15: Sociodemographic characteristics of participants in the partisan target study in comparison to sociodemographic characteristics of participants in the 2020 wave of the American National Election Studies (ANES)

Variable	N	Percent	N	Percent	Dif. Percent
	(Partisan Study)	(Partisan Study)	(ANES 2020)	(ANES 2020)	(Partisan-ANES)
Gender	3734		8226		
... Female	1843	49.4%	4262	52%	-2.6%
... Male	1803	48.3%	3964	48%	0.3%
Age Group	3732		7951		
... Born 1991 or later	1729	46.3%	1484	19%	27.3%
... Born 1975-1990	1292	34.6%	2076	26%	24.2%
... Born 1959-1974	529	14.2%	2186	27%	-12.8%
... Born 1943-1958	179	4.8%	1792	23%	-18.2%
... Born 1927-1942	3	0.1%	404	5%	-4.9%
... Born 1911-1926	0	0%	8	0%	0%
Race and Ethnicity	3709		8198		
... Black	324	8.7%	935	11%	-2.3%
... Hispanic	306	8.3%	1108	14%	-5.7%
... Race other/multiple	406	10.9%	773	9%	1.9%
... White	2673	72.1%	5383	66%	6.1%
Political Identity	3734		8251		
... Democrat	2073	55.5%	3808	46%	9.5%
... Independent	967	25.9%	976	12%	13.9%
... Republican	694	18.6%	3467	42%	-23.4%
Education	3734				
... College	1831	49%			
... High school graduate	1094	29.3%			
... Less than high school	37	1%			
... PhD	88	2.4%			
... Postgraduate (e.g. Masters)	480	12.9%			
... Professional degree	164	4.4%			
Education (ANES)			8147		
... Less than high school graduate			98	1%	
... High School (Grades 9-12)			2687	33%	
... Some College, no Degree			2376	29%	
... College Degree/ Post-grad			2986	37%	

Table S16: Logit models underlying the marginal effects displayed in Figure 5. Cell entries are regression coefficients (not exponentiated) and standard errors are shown in parentheses.

	Dependent variable:
	Any Moderation
	Pooled results
Uncivil post	1.834^{***}
	(0.165)
Intolerant post	1.806^{***}
	(0.166)
Threatening post	3.103^{***}
	(0.163)
No group mentioned (movie only)	-1.342^{**}
	(0.408)
Anti-target post	(ref. group)
Respondent PID: Democrat	0.316^{**}
	(0.100)
Respondent PID: Republican	-0.081
	(0.130)
Respondent PID: Independent	(ref. group)
Study: Democrats	0.315^{***}
	(0.083)
Study: Republicans	(ref. group)
(Intercept)	-3.078^{***}
	(0.170)
Num.Obs.	3,734
Log Likelihood	−1,758.435
Akaike Inf. Crit.	3,532.869

Note: *p<0.05; **p<0.01; ***p<0.001. The dependent variable “Any Moderation” takes the value of 1 if participants indicate support for any form of moderation, and 0 if participants responded “Leave it, do nothing”.

Table S17: Logit models underlying the contrasts displayed in Figure 5. Cell entries are coefficients (not exponentiated, left column) and standard errors (right column).

	Dependent variable:	
	Any Moderation	
	Coef. estimate	Std. error
(Intercept)	-4.644	1.005
No group mentioned (movie only)	0.501	1.423
Uncivil post	3.615	1.031
Intolerant post	3.455	1.028
Threatening post	4.409	1.023
Republican Respondent	2.636	1.055
Democrat Respondent	1.396	1.067
Study: Democrats	1.674	1.088
No group mentioned (movie only) x Republican Respondent	-1.355	1.630
Uncivil post x Republican Respondent	-2.609	1.109
Intolerant post x Republican Respondent	-2.596	1.108
Threatening post x Republican Respondent	-2.911	1.099
No group mentioned (movie only) x Democrat Respondent	-2.024	1.779
Uncivil post x Democrat Respondent	-1.494	1.103
Intolerant post x Democrat Respondent	-1.336	1.100
Threatening post x Democrat Respondent	-0.486	1.093
No group mentioned (movie only) x Study: Democrats	-1.501	1.794
Uncivil post x Study: Democrats	-1.181	1.132
Intolerant post x Study: Democrats	-1.431	1.133
Threatening post x Study: Democrats	-1.216	1.123
Republican Respondent x Study: Democrats	-1.687	1.189
Democrat Respondent x Study: Democrats	-1.031	1.177
No group mentioned (movie only) x Republican Respondent x Study: Democrats	1.081	2.239
Uncivil post x Republican Respondent x Study: Democrats	1.485	1.277
Intolerant post x Republican Respondent x Study: Democrats	0.915	1.300
Threatening post x Republican Respondent x Study: Democrats	1.438	1.266
No group mentioned (movie only) x Democrat Respondent x Study: Democrats	0.866	2.332
Uncivil post x Democrat Respondent x Study: Democrats	0.793	1.235
Intolerant post x Democrat Respondent x Study: Democrats	1.445	1.236
Threatening post x Democrat Respondent x Study: Democrats	0.764	1.226
Num.Obs.	3734	
Log Likelihood	−1,724.581	
Akaike Inf. Crit.	3,509.162	

Note: Reference categories = Type of post: Anti-target post, Political Identity = Independent Respondent, Study = Study: Republicans. The dependent variable “Any Moderation” takes the value of 1 if participants indicate support for any form of moderation, and 0 if participants responded “Leave it, do nothing”.

Table S18: Estimates and 95% confidence intervals displayed in Figure 5 (top and bottom panel) - Study II (partisans)

	Dependent variable:	
	Any Moderation	
Variable	Estimate	CI
Causal effects ↓		
Threatening language	0.523	[0.486,0.56]
Uncivil language	0.225	[0.19,0.26]
Intolerant language	0.219	[0.185,0.254]
Anti-target (baseline)	(ref. group)	(ref. group)
No group mentioned	-0.044	[-0.065,-0.024]
Target: A Democrat	0.049	[0.024,0.075]
Target: A Republican (baseline)	(ref. group)	(ref. group)
Observables ↓		
Democratic respondent	0.05	[0.019,0.08]
Republican respondent	-0.012	[-0.05,0.026]
Independent respondent	(ref. group)	(ref. group)

			Dependent variable:	
			Any Moderation	
Treatment group	Target	Respondents PID	Estimate	CI
Effect of intolerant language	A Republican target	Among Republicans	0.122	[0.007,0.237]
Effect of intolerant language	A Democratic target	Among Republicans	0.040	[-0.071,0.152]
Effect of intolerant language	A Republican target	Among Democrats	0.207	[0.144,0.269]
Effect of intolerant language	A Democratic target	Among Democrats	0.315	[0.246,0.384]
Effect of threatening language	A Republican target	Among Republicans	0.257	[0.127,0.386]
Effect of threatening language	A Democratic target	Among Republicans	0.308	[0.178,0.438]
Effect of threatening language	A Republican target	Among Democrats	0.625	[0.560,0.690]
Effect of threatening language	A Democratic target	Among Democrats	0.635	[0.567,0.703]
Effect of uncivil language	A Republican target	Among Republicans	0.150	[0.034,0.266]
Effect of uncivil language	A Democratic target	Among Republicans	0.212	[0.084,0.340]
Effect of uncivil language	A Republican target	Among Democrats	0.207	[0.146,0.268]
Effect of uncivil language	A Democratic target	Among Democrats	0.226	[0.159,0.293]
Num. Obs.	3734			
Log Likelihood	−1,724.581			
Akaike Inf. Crit.	3,509.162			

Note: The top table corresponds to the estimates visualized in the top panel of Figure 5. The estimates appearing in the bottom panel of Figure 5 can be found in the lower part of the table and they correspond to treatment differences compared to the Anti-target group split by partisanship and treatment group. When accounting for multiple comparisons (using False Discovery Rate) partisan differences related to the intolerant and uncivil conditions cease to be statistically distinguishable. The dependent variable "Any Moderation" takes the value of 1 if participants indicate support for any form of moderation, and 0 if participants responded "Leave it, do nothing".

Table S19: Percentage of participants preferring any form of moderation by treatment groups - Study II (partisans)

group	Democrats (Target) (%)	Republicans (Target) (%)	Partisans (Pooled) (%)
all	27	22	27
Control A: No group mentioned	2	2	2
Control B: Anti-target	7	5	6
T1: uncivil post	32	25	29
T2: intolerant post	32	24	28
T3: threatening post	61	55	58
Observations (N)	2076	2078	3734

Note: We considered that participants prefer any form of moderation if they selected any of "Permanently remove the post", "Place a warning label on the post", "Reduce how many people can see the post", or "Suspend the person's account" as their preferred action against the shown post. The "Control A: No group mentioned" is fully included in the Democrats' target study and the Republicans' target study to have a similar number of participants in each experimental group.

Table S20: Preferences for the type of moderation by treatment groups for all experiments in Study II (partisans)

treatment	how to handle the post	Partisans (Pooled)	Democrats (Target)	Republicans (Target)
		(%)	(%)	(%)
Control A: No group mentioned	Leave it, do nothing	98.3	98.3	98.3
Control A: No group mentioned	Permanently remove the post	0.2	0.2	0.2
Control A: No group mentioned	Place a warning label on the post	0.9	0.9	0.9
Control A: No group mentioned	Reduce how many people can see the post	0.5	0.5	0.5
Control B: Anti-target	Leave it, do nothing	94.0	92.8	95.2
Control B: Anti-target	Permanently remove the post	1.0	1.0	1.0
Control B: Anti-target	Place a warning label on the post	3.1	4.1	2.2
Control B: Anti-target	Reduce how many people can see the post	1.8	1.9	1.7
Control B: Anti-target	Suspend the person's account	0.1	0.2	
T1: uncivil post	Leave it, do nothing	71.3	68.0	74.6
T1: uncivil post	Permanently remove the post	5.1	7.5	2.7
T1: uncivil post	Place a warning label on the post	17.9	19.1	16.6
T1: uncivil post	Reduce how many people can see the post	4.6	4.4	4.9
T1: uncivil post	Suspend the person's account	1.1	1.0	1.2
T2: intolerant post	Leave it, do nothing	71.9	67.8	76.0
T2: intolerant post	Permanently remove the post	5.1	6.8	3.4
T2: intolerant post	Place a warning label on the post	17.2	17.9	16.6
T2: intolerant post	Reduce how many people can see the post	4.2	5.1	3.4
T2: intolerant post	Suspend the person's account	1.6	2.4	0.7
T3: threatening post	Leave it, do nothing	41.6	38.6	44.5
T3: threatening post	Permanently remove the post	17.1	17.7	16.4
T3: threatening post	Place a warning label on the post	26.1	24.8	27.4
T3: threatening post	Reduce how many people can see the post	7.9	9.2	6.7
T3: threatening post	Suspend the person's account	7.3	9.7	5.0
Observations (N)		3734	2076	2078

Note: 1. The "Control A: No group mentioned" is common for both targets (Democrats and Republicans), 2. Some types of moderation received no support from respondents in some of our groups (e.g."Control A: No group mentioned").

Content moderation practices and public opinion in the US and beyond

In the complex landscape of platform regulation, legislators in various countries have grappled with the responsibilities of online platforms for user-generated content and the regulation of online content. The United States has placed particular importance on freedom of speech values and this is reflected in its approach to regulating online content (Gillespie, 2018; Kohl, 2022). Section 230 of the 1996 U.S. Telecommunications Act has created a safe harbor for online platforms, as it (i) guarantees that platforms are not considered publishers and are therefore not to be held liable for any content posted by their users, while at the same time, (ii) it allows them to moderate their platforms by deleting posts without turning them into publishers and making them liable for future content (Gillespie, 2018, 30). As elaborated by Gillespie, such safe harbors are very advantageous from a legal perspective (Gillespie, 2018, 31), and platforms have strong motivations to "hold on to the safe harbor protections enshrined in Section 230, shielding them from liability for nearly anything that their users might say or do." (Gillespie, 2018, 34).

In examining platform governance beyond the U.S., it is clear that other nations adopt differing approaches to content moderation and regulation. European countries have placed a greater emphasis on combating harmful speech and hate speech than the U.S. The European approach, as highlighted by Kohl (2022), puts more emphasis on removing hate and harmful speech from the public domain and protecting the equality and inherent dignity of citizens in the public sphere, while the American approach prioritizes protecting the First Amendment's guarantee of freedom and keeping government interference in the online space minimal (Kohl, 2022). For instance, in Germany, the Network Enforcement Act (NetzDG) has mandated major platforms remove or block access to obviously illegal content within 24 hours after receipt of a complaint - with penalties up to 50 million Euro (e.g., Gorwa, 2021; Heldt, 2019). Additionally, platforms are obligated to create biannual transparency reports on their moderation activities if they receive more than 100 complaints per calendar year (Heldt, 2019) (for more insights on the regulatory developments in the case of the German NetzDG, see Gorwa (2021)). Gillespie (2018, 36–39) outlines what different countries consider illegal content. For example, in France and Germany, laws prohibit the promotion of "Nazism, anti-Semitism, and white supremacy," while Argentina's anti-discrimination law prohibits discriminatory or racist content. These laws enable them to hold platforms accountable and force them to remove such content. However, some countries (e.g., China, Egypt, Iran, Pakistan, Tunisia, and the United Arab Emirates) have enacted laws that criminalize speech that criticizes the government or public order. These laws raise questions about over-filtering as they not only enable the silencing of political activists but also allow for the blocking of entire pages, sites, or platforms. While the United States offers the aforementioned safe harbor for platforms as they are not responsible for user-generated content on their platform under Section 230 legislation, a law in Russia from 2009 holds website owners accountable for users' posts and comments, and allows the government to force the removal of politically undesirable material (Gillespie, 2018, 38).

While several countries have implemented regulations, notable recent regulatory efforts have occurred at a supranational level (for a deeper analysis of governance within the EU context, see, e.g., Busch (2022) and Mügge (2023)). The Digital Service Act (DSA) in Europe establishes various obligations on online platforms, such as publishing transparency reports on any content moderation employed by the platform (see e.g., Busch, 2022, 61–62; European Parliament, 2022). Furthermore, following the implementation of the DSA, EU member states can require platforms to delete posts that violate national laws. The content moderation ecosystem is undergoing changes, and research suggests the EU is adapting the existing regulatory framework, in response to the shift from content moderation to infrastructure moderation (Busch, 2022). An example of infrastructure moderation—a form of meta-moderation at higher levels—is the removal of apps from an App Store, as was the case with Parler from the Android and iOS App Store following the January 6 United States Capitol attack (Busch, 2022).

However, besides national and supranational legislators, social media firms also have interests in moderating content. Social media platforms implement their own policies and rules, and choose the circumstances under which they moderate content, particularly when there are economic incentives to do so, such as maintaining or increasing advertising revenue (Gillespie, 2018, 34–35; Klonick, 2017, 1627–1630). Alongside this major motivating driver, pursuing corporate responsibility can also be a motivating factor (Klonick, 2017, 1625–1627). Content moderation has always been a fundamental aspect of online platforms (Gillespie, 2018); as highlighted by Brunton (2013), spam is just one example of the necessity for basic content moderation. In fact, a wide variety of methods and strategies are used by social media firms for self-governance. Common approaches go beyond content removal and the suspension of users and encompass strategies such as automated moderation, so-called algorithmic moderation systems, that change the sorting of content or its visibility on the platform, for instance, depending on toxicity classifiers, making some content less visible than other content – instead of removing it (Gillespie, 2022; Gorwa, Binns and Katzenbach, 2020). However, the scale of some of the challenges may vary greatly depending on the technical solution employed by the platforms. In some instances, such as with algorithmic moderation systems, it can even exacerbate issues like a lack of transparency, or challenges surrounding fairness and equality (Gorwa, Binns and Katzenbach, 2020).

When it comes to public opinion regarding content moderation, insights from cross-country surveys might give insights into the importance of country context. Particularly including countries where the legal context in relation to the protection of vulnerable groups, the preservation of equality, political censorship and freedom of speech differs from that of the US could yield important new, and possibly unexpected insights. This is because while, on the one hand, there is a severe lack of comparative evidence on freedom of speech attitudes worldwide, the few cross-national public opinion surveys that exist show that publics in European, Asian and Latin American countries are not very far from the US when it comes to freedom of speech attitudes or to tolerance of offensiveness (Pew Research Center, 2015). For example, while in a global Pew Research survey 71% of Americans reported that "people can say what they want", that percentage was not too far from that of citizens in Latin America (69%) or Europe (65%). Similarly, while in a different question in the same survey 67% of Americans reported that they believe "people should be able to make statements that are offensive to minority groups publicly", this statement was also supported by majorities in countries like Spain (57%), the UK (54%), Australia (56%) and Mexico (65%). There are exceptions to this rule (e.g., Germany (27%), Italy (32%), South Korea (42%), Argentina (49%), and Brazil (48%)). In all, given the shortage of comparative empirical evidence, while we acknowledge that the free speech framework in the US makes our work a special case, we also note the necessity for more cross-national research as this value is clearly cherished among majorities in other countries.

Further analysis of incivility and intolerance as distinct constructs

To further analyze whether respondents distinguish between uncivil and intolerant posts, we tested whether they elicit different emotional reactions. We do find that intolerance induces significantly more negative emotions than incivility across the board (see Table S21) and conclude that these two categories are indeed perceived differently by respondents.

Table S21: Multiple linear regression on participants' emotions anger and disgust measured with slider questions [0,100]

	Dependent variable:	
	Anger (Pooled Study I&II)	Disgust (Pooled Study I&II)
Anti-target	15.992^{***}	21.316^{***}
	[13.961, 18.022]	[19.070, 23.562]
Uncivil post	27.833^{***}	37.230^{***}
	[25.797, 29.868]	[34.978, 39.482]
Intolerant post	32.329^{***}	42.191^{***}
	[30.296, 34.362]	[39.942, 44.439]
Threatening post	40.152^{***}	51.563^{***}
	[38.122, 42.183]	[49.317, 53.809]
Constant	4.273^{***}	4.862^{***}
	[2.750, 5.797]	[3.177, 6.547]
Observations	8,864	8,864
Adjusted R^2	0.182	0.238

Note: The table shows coefficients of estimated effects of each post type, relative to the reference category, on the respective emotion and 95% confidence intervals in square brackets. *p<0.05; **p<0.01; ***p<0.001, Reference category: No group mentioned.

Further analysis of heterogeneous effects

As moderation choices would assume a level of platform usage and familiarity, we further analyzed whether treatment effects vary with different levels of social media usage. We find no substantive differences in our treatment effects across different levels of social media platform usage (see Table S22).

Table S22: Logistic regression predicting support for any form of content moderation with treatment groups and social media usage.

	Dependent variable:
	Any Moderation
Anti-target	2.066*
	(1.046)
Uncivil post	3.548***
	(1.016)
Intolerant post	3.429***
	(1.020)
Threatening post	4.765***
	(1.013)
Frequent social media user	0.738
	(1.020)
Anti-target x Frequent social media user	−0.407
	(1.067)
Uncivil post x Frequent social media user	−0.466
	(1.036)
Intolerant post x Frequent social media user	−0.366
	(1.039)
Threatening post x Frequent social media user	−0.628
	(1.033)
Constant	−4.625***
	(1.002)
Observations	8,864
Log Likelihood	−1,771.050
Akaike Inf. Crit.	3,562.100

Note: The table shows coefficients (not exponentiated) of estimated effects. We consider frequent social media users those who visit social media platforms “Every day” or “At least once a week but not every day”, and infrequent social media users those who responded “A few times a month” or “Less often”. *p<0.05; **p<0.01; ***p<0.001, Reference categories = Type of post: No group mentioned, Usage of social media: Infrequent social media user. The dependent variable “Any Moderation” takes the value of 1 if participants indicate support for any form of moderation, and 0 if participants responded “Leave it, do nothing”.

References

Brunton, Finn. 2013. *Spam: A shadow history of the Internet.* Mit Press.

Busch, Christoph. 2022. "Regulating the Expanding Content Moderation Universe: A European Perspective on Infrastructure Moderation." *UCLA Journal of Law & Technology* 27:32–79.

European Parliament. 2022. "Digital Services Act: Agreement for a transparent and safe online environment.".
URL: *https://www.europarl.europa.eu/news/en/press-room/20220412IPR27111/digital-services-act-agreement-for-a-transparent-and-safe-online-environment*

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

Gillespie, Tarleton. 2022. "Do Not Recommend? Reduction as a Form of Content Moderation." *Social Media + Society* 8(3):1–13.

Gorwa, Robert. 2021. "Elections, institutions, and the regulatory politics of platform governance: The case of the German NetzDG." *Telecommunications Policy* 45(6):102145.

Gorwa, Robert, Reuben Binns and Christian Katzenbach. 2020. "Algorithmic content moderation: Technical and political challenges in the automation of platform governance." *Big Data & Society* 7(1).

Heldt, Amélie Pia. 2019. "Reading between the lines and the numbers: An analysis of the first NetzDG reports." *Internet Policy Review* 8(2).

Klonick, Kate. 2017. "The new governors: The people, rules, and processes governing online speech." *Harvard Law Review* 131:1598–1670.

Kohl, Uta. 2022. "Platform regulation of hate speech–a transatlantic speech compromise?" *Journal of Media Law* 14(1):25–49.

Mügge, Daniel. 2023. "The securitization of the EU's digital tech regulation." *Journal of European Public Policy* 30(7):1431–1446.

Pew Research Center. 2015. "Global support for principle of free expression, but opposition to some forms of speech." Pew Research Center.