Journal of Experimental Social Psychology 91 (2020) 104031

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Censoring political opposition online: Who does it and why

Ashwini Ashokkumar[a,*], Sanaz Talaifar[a], William T. Fraser[a], Rodrigo Landabur[b], Michael Buhrmester[c], Ángel Gómez[d], Borja Paredes[e], William B. Swann Jr[a]

[a] Department of Psychology, University of Texas at Austin, 108 E. Dean Keeton, Austin, TX 78712-0187, United States
[b] Department of Psychology, Universidad de Chile, Ignacio Carrera Pinto 1045, Nuñoa, Región Metropolitana, Chile
[c] Institute of Cognitive and Evolutionary Anthropology, University of Oxford, 51-53 Banbury Road, Oxford OX2 6PE, UK
[d] Department of Social and Organizational Psychology, Facultad de Psicología (UNED), C/Juan del Rosal, 10, 28040 Madrid, Spain
[e] Department of Communication Theories and Analysis, Facultad de Ciencias de la Información (Universidad Complutense de Madrid), Avenida Complutense, 3, 28040 Madrid, Spain

ARTICLE INFO

☆This paper has been recommended for acceptance by Ashwini Ashokkumar

Keywords:
Censorship
Selective censoring
Identity politics
Moderators
Identity fusion
Social media

ABSTRACT

As ordinary citizens increasingly moderate online forums, blogs, and their own social media feeds, a new type of censoring has emerged wherein people selectively remove opposing political viewpoints from online contexts. In three studies of behavior on putative online forums, supporters of a political cause (e.g., abortion or gun rights) preferentially censored comments that opposed their cause. The tendency to selectively censor cause-incongruent online content was amplified among people whose cause-related beliefs were deeply rooted in or "fused with" their identities. Moreover, six additional identity-related measures also amplified the selective censoring effect. Finally, selective censoring emerged even when opposing comments were inoffensive and courteous. We suggest that because online censorship enacted by moderators can skew online content consumed by millions of users, it can systematically disrupt democratic dialogue and subvert social harmony.

1. Introduction

In the run-up to the 2016 presidential elections, the moderators of a large online community of Trump supporters deleted the accounts of over 2000 Trump critics. The moderators even threatened to "throw anyone over our walls who fails to behave themselves" (Conditt, 2016). This phenomenon of silencing challenging voices on social media is not limited to the hundreds of thousands of designated moderators of online communities and forums; even ordinary citizens can delete comments on their own posts and report or block political opponents (Linder, 2016). To study this new form of censorship, we developed a novel experimental paradigm that assessed the tendency for moderators to selectively censor (a) content that is incongruent with their political cause (a political position or principle that people strongly advocate) and (b) the authors of such incongruent content. The studies also tested whether identity-related processes amplified the selective censorship of cause-incongruent content. Further, we tested whether the identity-driven selective censoring of political opponents' posts occurs even when opponents express their views in a courteous and inoffensive manner. To set the stage for this research, we begin with a discussion of past literature on biased exposure to online content.

1.1. Biased exposure to online content: selective information-seeking and avoidance

Behavioral scientists have long noted that people create social environments that support their values and beliefs (McPherson et al., 2001). People gravitate to regions, neighborhoods or occupations in which they are surrounded by individuals with similar personalities (Rentfrow et al., 2008) or political ideologies (Motyl et al., 2014). Once in these congruent environments, people are systematically exposed to information that aligns with their own views (Hart et al., 2009; Sears and Freedman, 1967). In addition, people actively display biases in behavior (e.g. choice of relationship partners) and cognition (e.g. attention, recall, and interpretation of feedback) that encourage them to see more support for their beliefs than is justified by objective reality (Garrett, 2008).

Parallel phenomena can occur in virtual worlds. People often find themselves in online bubbles of individuals who share political beliefs and information with each other but not with outsiders (Adamic and Glance, 2005; Barberá et al., 2015). They also actively seek websites or online communities that support their pre-existing opinions (Garimella and Weber, 2017; Iyengar and Hahn, 2009), and follow or connect with individuals whose opinions they endorse (Bakshy et al., 2015; Brady

* Corresponding author at: Department of Psychology, University of Texas at Austin, 108 E. Dean Keeton, Austin, TX 78712-0187, United States.
E-mail address: ashwinia@utexas.edu (A. Ashokkumar).

https://doi.org/10.1016/j.jesp.2020.104031
Received 15 December 2019; Received in revised form 15 July 2020; Accepted 16 July 2020

et al., 2017). And when they process information that they encounter, they display confirmation biases that warp their visions of reality (Hart et al., 2009; Van Bavel and Pereira, 2018). Some evidence also suggests that in addition to actively seeking attitude-consistent online content, people also avoid attitude-inconsistent content (Garrett, 2009a). Importantly, biases in information seeking are strongest for content related to political and moral issues (Stroud, 2017) and are most prevalent among those who have strong views or ideologies (Boutyline and Willer, 2017; Hart et al., 2009; Lawrence et al., 2010).

Although researchers have investigated biases in how people *seek*, *consume*, or *avoid* information in online contexts, to the best of our knowledge they have yet to examine how people might *influence* the content to which they and others are exposed through censorship. It is increasingly possible for individuals to censor others in online contexts by deleting others' comments on their own posts and pages (John and Dvir-Gvirsman, 2015; Sibona, 2014). For moderators of popular social media pages and large forums, the scope of their ability to censor is multiplied as they often exercise control over content that millions view (Matias, 2016a; Wright, 2006).

Censorship is more extreme than biased information seeking because, in addition to biasing one's own online environment, censorship delimits the online content that other people are exposed to. Also, by silencing dissenters, censorship prevents them from voicing their views. And although the psychological processes underlying censorship may overlap with some of the defensive motivations producing selective information seeking (Hart et al., 2009), censorship may in addition entail a hostile motivation to nullify opponents of the cause.

1.2. Censorship in offline and online environments

The majority of past studies on censorship have examined the association between political orientation and attitudes toward censorship. Whereas some studies have suggested that conservatives support censorship (Fisher et al., 1999; Hense and Wright, 1992; Lindner and Nosek, 2009), others have reported evidence of censorship by people on both sides of the political spectrum (Crawford and Pilanski, 2014; Suedfeld et al., 1994). One limitation of this work is that researchers have typically explored people's *attitudes* toward censorship rather than their censoring *behaviors*. Further, to our knowledge, no studies have systematically examined censoring behaviors in online settings.

As public pages and forums are increasingly moderated by everyday citizens (Matias, 2016a), the power to censor others is now widely available. For example, on the popular social media platform Reddit, almost 100,000 community moderators have the power to delete comments or entirely ban accounts associated with millions of users (https://mods.reddithelp.com/). Even internet users who have no particular stature within online communities are able to moderate other people's comments on their own posts and blogs. People can "report" social media posts they find disagreeable (John and Dvir-Gvirsman, 2015; Sibona, 2014) or simply delete or hide cause-incongruent comments on their own posts or blogs. Given that censoring in online contexts is easier (e.g., requires a single click) and may have fewer personal repercussions relative to offline contexts (e.g., more anonymity), it seems likely that online censoring will become increasingly prevalent. Here, we examine people's tendency to selectively censor content that is incongruent with a political cause they support.

1.3. Identity as a censorship amplifier

Not everyone will be equally motivated to selectively censor cause-incongruent content. For example, motivation to censor content will be particularly high when it challenges a political cause with which people's identities are strongly "fused" (Swann Jr et al., 2012). For people who are strongly fused with a cause, threats to the cause will feel like threats to the self. This will induce strongly fused people to be particularly reactive to threatening content (Gómez et al., 2011; Swann Jr et al., 2009). They may, for instance, go to great lengths to protect their group (Swann Jr et al., 2014) and are even attempt to inflict serious harm on threatening outgroups (Fredman et al., 2017). Therefore, we expect that strongly fused individuals would be especially apt to selectively censor incongruent content to preserve their cause against challenges.[1]

Although we focused primarily on identity fusion as a potential amplifier of censorship, we also investigated several other identity-related measures that have been associated with intolerance of political opposition. The literature on self and identity broadly suggests that people's social identities relating to political groups and causes are potent predictors of action intended to advance one's group or cause (e.g., Ashokkumar et al., 2019; Swann Jr et al., 2012; Tajfel and Turner, 1979) and counter opponents (Brewer, 2001; Fredman et al., 2017). In line with this reasoning, we investigated the effects of various other identity-related measures: indices of attitude strength, moral conviction, and identification with other supporters of the cause. Attitude strength and moral conviction are part of people's identities because their preferences and moral values are important parts of their self-related mental representations (McAdams, 1995). Past research on attitude strength has revealed that people who hold extreme views about a cause or whose views are associated with feelings of certainty and personal significance are intolerant of others with dissimilar attitudes (e.g., Singh and Ho, 2000; Singh and Teoh, 1999). Similarly, moral convictions reflect people's deeply held beliefs regarding the morality of a cause (Skitka and Mullen, 2002) and is known to predict an aversion to attitudinally dissimilar others (Skitka et al., 2005). Finally, we assessed participants' identification with cause supporters, since identification has been found to be a potent predictor of pro-cause action (Thomas et al., 2016). Although the foregoing variables have all been associated with intolerance of outgroups and are important components of people's identities (i.e. their mental self-representations), the causal, structural, and temporal relationships between these variables have not been clearly established. For example, it is unclear whether strong moral convictions cause greater group identification or the reverse (Van Zomeren et al., 2012; Zaal et al., 2017). Similarly, the temporal relationship between fusion with cause and group identification is not clear (Gómez et al., 2019). Prior work has shown that identity fusion is associated with moralized attitudes (Talaifar and Swann Jr, 2019) but the causal relationship between these variables is unclear. Nevertheless, given that these variables have been found to predict a suite of behaviors related to intolerance of political opposition, we included them as potential predictors of selective censoring.

1.4. Overview of studies

The current research had two primary goals. First, we asked whether people assigned to moderate online content would selectively censor opposition to their political causes by deleting opposing comments and banning opponents from a forum. Second, we examined whether people whose cause-related beliefs were rooted in their identities would be especially likely to selectively censor incongruent content. In all studies, we recruited participants from the United States (US). Based on past reports that biases in information consumption are stronger for political and moral issues (Stroud, 2017), we focused on political causes that are deemed to have a moral component. Specifically, we chose abortion rights (Studies 1–2) and gun rights (Study 3) as the focal issues. We also selected these issues because they are highly controversial in the US to raise the likelihood that most people would have relevant opinions. In fact, many believe that over the last half

[1] Selective censorship can occur as a result of two processes: *greater* censoring of cause-incongruent content and/or *less* censoring of cause-congruent content. We did not have an a priori hypothesis regarding which of these selective censoring processes fusion would amplify.

century these issues determined the outcome of multiple elections in the U.S. (Leber, c., 2016; Riffkin, 2015).

All studies used a longitudinal design in which we measured all predictors at Time 1 (T1) and censoring at Time 2 (T2). At T1, we measured participants' position on an issue (e.g., abortion rights) and their identity fusion with the corresponding cause (e.g., pro-life or pro-choice cause). In Studies 2 and 3, we also measured other prominent identity-related measures, including strength of attitudes, moral conviction, and identification with cause supporters. As part of a seemingly unrelated study administered two weeks later (Time 2 or "T2"), we measured participants' censoring behavior using a novel simulation of an online forum. We sought participants' assistance in moderating the content of a putative online forum. Participants read comments and decided whether the comments needed to be retained or removed from the forum. Comments they chose to remove were considered "censored." Each comment was systematically manipulated to be either congruent or incongruent with the participant's cause and either offensive or inoffensive. In Studies 2 and 3, we also asked participants whether the authors of the congruent and incongruent comments they read should be banned from the forum.

We operationalized selective censorship as either a preference for cause-congruent content or an intolerance of cause-incongruent content. We expected that cause supporters would selectively censor comments incongruent with their cause (Studies 1–3) and selectively ban the author of those incongruent comments (Study 2 & 3). We also expected that people whose identities were strongly aligned ("fused") with the cause would be particularly likely to selectively censor incongruent comments (Studies 1–3) and selectively ban the authors of those comments (Study 2–3). We examined whether the effect of fusion was influenced by the presence of offensive language in the comments (Studies 1–3) and also whether the effect generalized to an array of other identity-related measures (Study 2 & 3). Further, in SOM-III we explored one potential mechanism driving the effect of fusion on selective censoring: strongly fused people's tendency to essentialize the cause. In all studies, we examined whether there were partisan differences in selective censoring (i.e. if selective censoring was stronger among pro-life vs. pro-choice supporters in Studies 1 and 2; pro-gun-rights vs. pro-gun-control supporters in Study 3), and we report any asymmetries between the two sides. For all three studies, we report all measures, manipulations, and exclusions.

2. Study 1

2.1. Study 1 method

2.1.1. Time 1 (T1)

2.1.1.1. Participants. In August 2017, we recruited 477 participants from Amazon's Mechanical Turk (MTurk), an appropriate source of data for our purposes given that MTurkers routinely review comments by actual website moderators (Schmidt, 2015).[2] Participants first indicated their position on the issue of abortion rights (pro-choice vs. pro-life vs. neither/don't know). Thirty-five participants who reported neutral or no views on abortion rights were not allowed to proceed because a person's pre-existing position on abortion rights needs to be known in order to identify which comments are congruent vs. incongruent with their cause. We removed 32 respondents with identical IP addresses or MTurk Worker IDs to eliminate the possibility of a single respondent completing the survey twice. We excluded four participants who failed our attention check (see SOM-I). Our final T1 sample had 406 participants (49.8% female; 71.6% White; M_{age} = 36.06; SD_{age} = 11.59; 274 pro-choice and 132 pro-life participants). The higher proportion of pro-choice participants is typical in liberal-skewed online crowdsourcing platforms such as MTurk (e.g., Ashokkumar et al., 2019). In this and all studies, sample size was determined prior to data analysis.

2.1.1.2. Identity measures. Participants completed the seven-item verbal fusion scale (α = 0.91, 95% CI = [0.89, 0.93]) measuring fusion with their cause (e.g. "I am one with the pro-life/pro-choice position"; Gómez et al., 2011). They also completed a five-item measure of the mediating mechanism explored in SOM-III: essentialist beliefs relating to the cause (α = 0.91, 95% CI = [0.90, 0.93]) adapted from Bastian and Haslam (2006); (e.g., "There are two types of people in this world: pro-life and pro-choice"). Both constructs were rated on seven-point scale ranging from 1 (*Strongly Disagree*) to 7 (*Strongly Agree*). We standardized the fusion and essentialism scores prior to analysis. Means, standard deviations, and inter-variable correlations in the final sample are reported in Table 1.

Participants provided demographic information before completing the survey (see https://osf.io/4jtwk/?view_only=10627a9892464e5aa90fe92360b846ad for a full list of measures). At the end of the study, participants learned that they might be contacted again for other studies. We did not specify when or why we would re-contact them because we wanted to discourage them from associating the first session of the study with the second.

2.1.2. Time 2 (T2)

2.1.2.1. Participants. Two weeks later we re-contacted the participants regarding a seemingly unrelated "comment moderation task." A total of 251 participants completed the second session of the study, amounting to a 38.2% attrition rate, which is comparable to previously reported attrition rates on MTurk (Stoycheff, 2016). There were no differences in fusion ($t(400) = -0.19$, $p = .85$, $d = -0.02$) between those who did vs. did not complete the second session of the study. We excluded 25 respondents with identical IP addresses or MTurk worker IDs and three participants who evaluated fewer than 50% of the comments in the comment moderation task, resulting in a final sample of 223 participants (52% female; 71.8% White; M_{age} = 38.36; SD_{age} = 11.99; 148 pro-choice and 75 pro-life participants) who completed both time points. We were unable to conduct an a priori power analysis because the lack of previous research on censoring made it difficult for us to estimate expected path coefficients, which is required for power analyses for Structural Equation Models (SEM; Muthén and Muthén, 2012). To give a general sense of how much power we had with the present sample size, we conducted a sensitivity analysis, which revealed that the sample had 80% power to detect a minimum effect size of f^2 = 0.04 in a multiple regression.

2.1.2.2. Comment moderation procedure. In the comment moderation task, participants read about a new blog purportedly launched with the goal of "encouraging discussion about current issues." We informed participants that we had received complaints regarding a surge in inappropriate comments posted on the blog and that we needed their help in deleting inappropriate comments. To make sure that participants took the task seriously, we informed them that the blog's administrator would delete all comments that they flagged. Participants then read a series of 40 statements that were adapted from comments from real online blogs and forums. Of the 40 comments, 15 were pro-choice (e.g.: "I love that even though Norma couldn't herself get an abortion (because of the terrible world we live in), she fought so hard to make sure other women could."), 15 were pro-life (e.g.: "I love that Lily didn't have an abortion even though she didn't want to be a parent. She hadn't planned a baby and wasn't ready for it, but she didn't get an abortion."), and 10 were irrelevant to the cause (e.g.: "I still can't wrap my head around this horrific, senseless act. Sickening."). Participants could recommend either deletion or retention of each comment. The full list of comments is available at https://osf.io/4jtwk/?view_only=10627a9892464e5aa90fe92360b846ad.

[2] Note that the data were collected before reports of drop in the quality of the MTurk participant pool surfaced in, 2018 (TurkPrime, 2018).

Table 1
Means, standard deviations, and correlations of measures in Study 1 (N = 223).

Variable	*M*	*SD*	1	2	3
1. Fusion with cause	4.71	1.39			
2. Censoring rate -congruent comments	0.20	0.19	−0.06		
3. Censoring rate - incongruent comments	0.26	0.22	0.13*	0.56**	
4. Censoring rate - irrelevant comments	0.34	0.16	0.12	0.56**	0.51**

Note. The censoring rates, ranging from 0 to 1, refer to the proportion of comments of each type (congruent, incongruent, or irrelevant) that participants censored. Fusion's effect on selective censoring is the difference between fusion's association with the censoring rates of congruent and incongruent comments. Fusion's effect was not influenced by position on abortion rights. * indicates $p < .05$. ** indicates $p < .01$.

For each participant, we calculated three censoring rates corresponding to the proportion of comments that the participant deleted among (a) congruent comments (i.e., comments endorsing the participant's position on abortion rights), (b) incongruent comments (i.e., comments against the participant's position on abortion rights), and (c) irrelevant comments (i.e., comments irrelevant to abortion rights). The three censoring rates were inter-correlated (see Table 1), which indicates that individual differences in people's general tendency to censor were relatively stable across comments.

2.1.2.3. Post-hoc assessment of comment offensiveness. To determine whether strongly fused people's tendency to selectively censor incongruent comments depended on whether the comments included offensive language, we asked five objective judges from MTurk to provide post-hoc ratings of each comment's offensiveness. Of the five judges, two were pro-choice, two were pro-life, and one was neutral (i.e., did not favor either side of the abortion debate). The judges were told that offensive comments were those that "a reasonable person would consider to be abusive, harassing, or involving hate speech or ad hominem attacks." The inter-judge reliability across the five judges was $\alpha = 0.84$. We coded each comment as offensive or inoffensive based on the judges' majority opinion (see SOM-I for more details). The offensive vs. inoffensive classification generated from the post-hoc pilot was then applied in the selective censoring analyses.[3] For each participant, we computed four censoring rates corresponding to the proportion of comments that the participant censored among comments of four categories: Offensive-Congruent, Offensive-Incongruent, Inoffensive-Congruent, and Inoffensive-Incongruent.

2.2. Study 1 results

2.2.1. Did people selectively censor comments incongruent with their cause?

To test whether people censored incongruent comments at a higher rate than congruent comments, censoring rates for incongruent vs. congruent comments were compared via a paired *t*-test. A significant effect emerged ($t(220) = 4.0, p < .001, d = 0.25$). On average, people censored 25.64% ($SD = 22.35$) of the incongruent comments they read but only 20.41% ($SD = 18.72$) of the congruent comments. Later in this section, we report differences in selective censoring between pro-life and pro-choice participants.

[3] When designing the Study 1 materials, we did not ensure that the three types of comments (i.e., pro-choice, pro-life, and irrelevant comments) were equally offensive. For example, the post-hoc offensiveness ratings suggest that the pro-life comments may have been generally less offensive than the pro-choice and irrelevant comments. For this reason, the estimates of censoring obtained in Study 2, in which we systematically varied offensiveness a priori, are more trustworthy.

2.2.2. Did identity fusion amplify the selectively censoring of incongruent comments?

We used structural equation modeling (SEM) for our analyses to simultaneously model fusion effects on two dependent variables: censoring rate for congruent and incongruent comments. We also conducted alternate analyses treating the difference between people's rates of censoring incongruent and congruent comments as the index of selective censoring and regressing the index over fusion (see SOM-II). Although this method feels intuitively appealing, it is not ideal because the method would not tell us whether any detected effect is driven by people's preference for congruent comments or their antagonism against incongruent comments. Past theorists have warned against conflating these two separate processes and recommend that each should be modeled separately (Garrett, 2009a, 2009b; Holbert et al., 2010). The SEM approach allows us to simultaneously model effects on censoring rates for congruent and incongruent comments treating them as two separate variables with different variances rather than assuming them to constitute a single variable. Note however that both the methods (SEM and computing a difference index) lead us to the same conclusions.

To evaluate our hypothesis that strongly fused people would be especially likely to selectively censor incongruent comments relative to congruent comments, we tested whether the effect of fusion on censoring incongruent comments (indicated by the c_1 path in Fig. 1) is significantly larger than the effect of fusion on censoring congruent comments (c_2 path). A significant difference between the two path coefficients (i.e., $\Delta c = c_1 - c_2$) would suggest that fusion is associated with disproportionately censoring incongruent, over congruent, comments. In this and all other models, we allowed for residual covariances between the censoring rates. In all the models, we used standardized scores for the continuous predictors, but we did not standardize the censoring rates (they ranged from 0 to 1) to allow the censoring effects to be interpreted in meaningful units. We report unstandardized regression coefficients.

Fusion was associated with censoring incongruent comments (c_1 path; $b = 0.03$, 95% CI = [0.001, 0.06], $p = .04$) but not with censoring congruent comments (c_2 path; $b = -0.01$, 95% CI = [−0.04, 0.01], $p = .38$). A Wald test revealed that the difference between the two paths was statistically significant, ($\chi^2(1) = 9.88, p = .002$), which is evidence for our main hypothesis that strongly fused individuals are more likely to selectively censor incongruent than congruent comments. To illustrate, participants who were strongly fused (1 *SD* above the mean) censored 29.56% of the incongruent comments they read but only 15.75% of the congruent comments, while those who were weakly fused (1 *SD* below the mean) did not censor incongruent comments (20.74%) any more than they censored congruent comments (20.37%). The significant c_1 path suggests that the effect of fusion on selective censoring is driven by strongly fused people's intolerance for incongruent comments rather than their leniency toward congruent comments.

Controlling for the censoring rate of comments irrelevant to abortion rights (to account for participants' general censoring rate and other response biases) did not alter the effect of fusion on selective censoring ($\chi^2(1) = 9.88, p = .002$). The fusion effect remained robust when we controlled for participants' position on abortion rights (i.e., pro-life vs. pro-choice; $\chi^2(1) = 8.33, p = .004$). Further, the fusion effect was not influenced by the participant's abortion rights position ($\chi^2(1) = 1.28$, $p = .26$), indicating that fusion was equally associated with selective censoring among both pro-life and pro-choice participants. In SOM-III, we report exploratory analyses testing whether essentialist beliefs about people's views on abortion rights mediates the fusion effect on selective censoring.

2.2.2.1. Did offensiveness influence the effect of fusion on selectively censoring?. We asked whether the tendency for strongly fused participants to selectively censor incongruent comments depended on

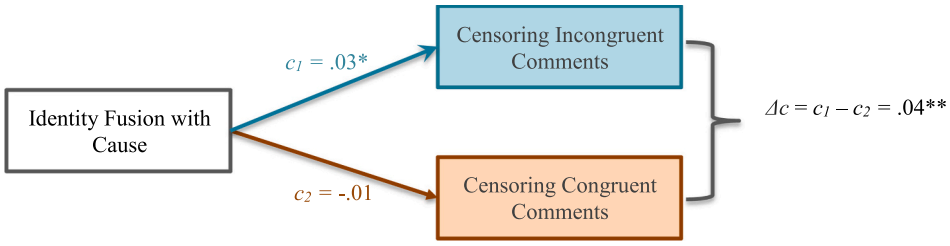

Fig. 1. Structural Equations Model depicting the effect of identity fusion on selective censoring of incongruent vs. congruent comments (Study 1). The c_1 and c_2 paths represent the effects of fusion on censoring incongruent and congruent comments respectively. The significant difference between the two paths (i.e., Δc) indicates that fusion is associated with selectively censoring incongruent comments. The coefficients reported are unstandardized. * indicates $p < .05$. ** indicates $p < .01$.

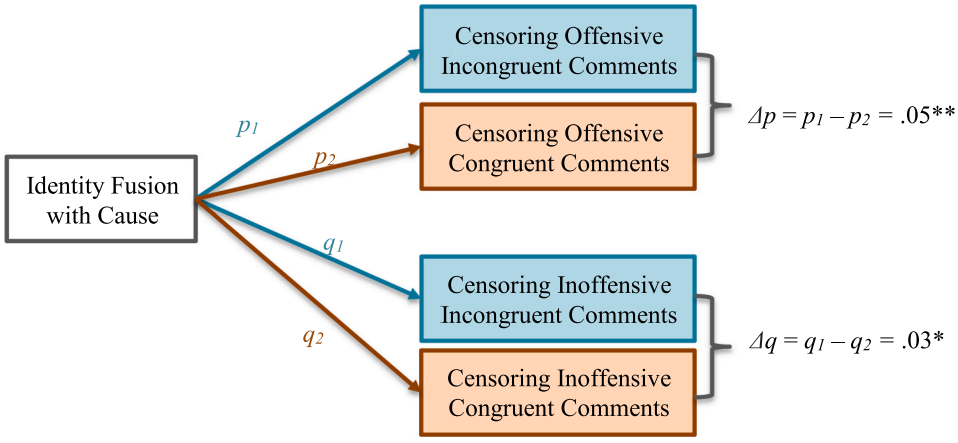

Fig. 2. Structural Equations Model examining the effect of identity fusion on selective censoring of incongruent vs. congruent comments among offensive and inoffensive comments (Study 1). Δp and Δq represent fusion's effects on selective censoring among offensive comments and inoffensive comments, respectively. The significant effects indicate that strongly fused people selectively censored incongruent comments whether the comments were offensive or inoffensive. See SOM-IV for path coefficients. * indicates $p < .05$. ** indicates $p < .01$.

how offensive the comments were. As depicted in Fig. 2, we modeled the paths from fusion to participants' censoring rates for four types of comments: Offensive-Congruent, Offensive-Incongruent, Inoffensive-Congruent, and Inoffensive-Incongruent. We allowed for residual covariances between the censoring rates.

We first computed the effects of fusion on selective censoring of incongruent vs. congruent comments separately for offensive and inoffensive comments. To compute the effect of fusion on selective censoring for offensive comments, we compared fusion's effect on censoring Offensive-Incongruent (path *p1*) vs. Offensive-Congruent (path *p2*) comments. The significant difference between the two *p* paths ($\Delta p = p1 - p2$, $b = 0.05$, 95% CI = [0.01, 0.09], $p = .008$) suggests that among offensive comments, strongly fused individuals selectively censored incongruent comments more than congruent comments. (Refer to SOM-IV for the path coefficients). Similarly, we computed fusion's effect on selective censoring for inoffensive comments as the difference between fusion's effect on censoring Inoffensive-Incongruent comments (path *q1*) vs. Inoffensive-Congruent comments (path *q2*). The resulting significant difference ($\Delta q = q1 - q2$; $b = 0.03$, 95% CI = [0.002, 0.05], $p = .04$) indicated that among inoffensive comments, participants censored incongruent comments more than congruent comments. In short, strongly fused individuals selectively censored incongruent comments more than congruent comments both when the comments were offensive and inoffensive.

Finally, to test whether strongly fused people's tendency to selectively censor incongruent comments was stronger for offensive comments, we compared the two selective censoring effects reported above for offensive vs. inoffensive comments. The difference ($\Delta p - \Delta q$) was non-significant ($\chi^2(1) = 2.10$, $p = .15$), suggesting that the effect of fusion on selective censoring was independent of the offensiveness of comments. That is, strongly fused individuals selectively censored incongruent, as opposed to congruent, comments regardless of whether the content of the comments included offensive language.

2.2.3. Did selective censoring of incongruent comments depend on people's ideologies?

Using a SEM model similar to the fusion analysis, we tested whether there were differences in people's tendency to selectively censor incongruent vs. congruent comments as a function of their stance on abortion rights (i.e., whether they were pro-choice or pro-life). Participants who endorsed the pro-life position showed a stronger tendency to selectively censoring incongruent comments relative to those who endorsed the pro-choice position ($\chi^2(1) = 7.36$, $p = .007$). Pro-life participants also reported marginally higher fusion levels than did pro-choice participants [$t(220) = 1.76$, $p = .08$, $d = 0.25$].

2.3. Study 1 discussion

Study 1 used a novel paradigm to explore people's censoring behaviors in online settings. People tended to censor online content more if the content was incongruent, rather than congruent, with their cause, and this tendency was higher among supporters of the pro-life cause. Importantly, identity-related processes amplified selective censoring of incongruent online content for people on both sides of the abortion rights cause. Specifically, the results showed that people whose identities were strongly fused with a cause were most willing to selectively censor online content posted by their ideological opponents. Interestingly, strongly fused people's tendency to selectively censor comments was driven by their intolerance for incongruent comments rather than an elevated affinity for congruent comments. Post-hoc analyses also showed that fusion's effect on selective censoring occurred regardless of whether the incongruent comments used offensive language. It is notable that strongly fused people showed a stronger selective censoring effect than weakly fused people even though they were not primed to think about their identity before reading the comments.

3. Study 2

Study 2 attempted to replicate Study 1 in a pre-registered longitudinal study. The method was largely similar to that of Study 1. To verify the preliminary findings from Study 1's post-hoc analysis on the effects of offensiveness, Study 2 systematically manipulated comment offensiveness a priori. The comments used in the study were pretested

and categorized as containing offensive vs. inoffensive content. This allowed us to more robustly probe whether the fusion effect on selective censoring was moderated by offensiveness. Further, it was not clear from Study 1 whether strongly fused people's tendency to selectively censor incongruent comments would extend to censoring the authors of the comments. To test this possibility, the study tested whether strongly fused individuals would opt to ban people who repeatedly posted content that threatened their position on the cause. The hypotheses were pre-registered prior to data collection (see https://osf.io/2jvau?view_only=754165d77cbe4e69baf6b11740b1a422).

Finally, although we have only focused on identity fusion thus far, we wanted to test whether the effects generalize to other identity-related measures explored in the broad literature: attitude strength, moral conviction, and identification with cause supporters. Studies have found that these constructs predict pro-cause action and an intolerance for opposition (e.g., Singh and Ho, 2000; Skitka et al., 2005; Thomas et al., 2016). We examined the extent to which each of these identity-related measures predicted selective censoring.

4. Study 2 method

4.1. Power analysis

An a priori power analysis was conducted using Monte Carlo simulations to estimate the sample size required to detect the SEM models reported in Study 1. As mentioned in our pre-registration, a sample of 345 participants was required to detect the selective censoring effect computed from the mediation model explored in Study 1 (see SOM-III) with an alpha of 0.05 and 80% power. In addition to replicating Study 1 effects, we wanted to test models examining the impact of the other identity-related measures (attitude strength, moral conviction, and identification with cause supporters) on censoring and also test a model with all identity-related measures simultaneously entered into a structural equation model. Because we had no easy way to estimate the path coefficients for these models, we estimated the required sample size by conducting a conservative power analysis using the models reported in Study 1. As mentioned in our pre-registration, we conducted Monte Carlo simulations to detect the Study 1 mediation model with a conservative alpha of 0.01 and found that we would need a sample size of 510. This conservative estimate would give us sufficient power to detect smaller effects than the ones reported in Study 1. Given the longitudinal nature of the study, we estimated that about 35% of the sample would either drop out between T1 and T2 or be excluded because of failing attention checks, and so we decided to recruit 800 participants at T1. The power analysis and exclusion criteria followed were specified in the pre-registration. Any deviations from the pre-registered plan are noted.

4.2. Comment offensiveness pretest

We wanted to systematically manipulate the offensiveness of comments. To classify comments as offensive vs. inoffensive, we conducted a pilot study on MTurk. We recruited five Mturkers who reported having neutral or no opinions about the abortion rights issue to be objective judges. We piloted 40 comments of which 20 were pro-choice and 20 were pro-life. For each position (pro-choice and pro-life), we piloted 10 comments that we believed contained offensive content and 10 that did not. The instructions provided to the objective judges were the same as in Study 1. The judges evaluated the content of each comment as either offensive or inoffensive. The inter-judge reliability across the five judges was $\alpha = 0.87$. For each of the four types of comments (Offensive-Prochoice, Inoffensive-Prochoice, Offensive-Prolife, and Inoffensive-Prolife), the seven comments with the highest levels of agreement among the judges were selected for the study. At least three of the five judges agreed on the categorization of the 28 comments that were finally selected for the study (see https://osf.io/4jtwk/?view_only=10627a9892464e5aa90fe92360b846ad for the final list of comments).

4.3. Time 1 (T1)

4.3.1. Participants

A sample of 793 participants from Prolific Academic completed the first part of the study in July 2019. The method followed was largely similar to Study 1. As mentioned in the pre-registration, only participants who endorsed either the pro-choice or pro-life position were eligible for the study. This was ensured by setting a pre-screening condition on Prolific such that the study posting was visible only to participants who had previously identified as pro-choice or pro-life. To be sure that the pre-screening worked, participants' views on abortion rights were measured again in the T1 survey, and 15 participants who indicated holding neutral views on abortion were excluded. We also excluded 29 participants who failed either of two attention checks or did not complete them (see SOM-I). Our final sample at T1 had 749 participants (48% female; 69.88% White; $M_{age} = 32.88$; $SD_{age} = 11.79$; 616 pro-choice and 133 pro-life participants).

4.3.2. Identity measures

As in Study 1, participants completed the seven items of the verbal fusion scale measuring fusion with their own position on the abortion rights (either pro-choice or pro-life) on a seven-point scale ($\alpha = 0.92$, 95% CI = [0.91, 0.93]). The survey also included measures of a series of identity-related measures including four facets of attitude strength such as attitude extremity ("What is your opinion about the pro-life/pro-choice position?"; 1 = *Strongly* against, 9 = *Strongly favor*; Binder et al., 2009), attitude centrality ("To what extent does your opinion toward the pro-life/pro-choice position reflect your core values and beliefs"; Clarkson et al., 2009), attitude certainty (e.g., "How certain are you of your opinion about the pro-life/pro-choice position?"; 1 = *Not certain at all*, 9 = *Extremely certain*; Fazio and Zanna, 1978) and attitude importance (e.g., "To what extent is the pro-life/pro-choice position personally important to you?"; Boninger et al., 1995; $\alpha = 0.91$, 95% CI = [0.89, 0.92]). Attitude extremity, centrality, and certainty were measured using one item each, and attitude importance was measured using two items. Attitude centrality and attitude importance used nine-point scales (e.g., 1 = *Not at all*; 9 = *Very Much*). We also measured moral conviction (e.g., "To what extent is your position on the pro-life position a reflection of your core moral beliefs and convictions?"; Skitka and Morgan, 2014) using two items on a five-point scale ($\alpha = 0.86$, 95% CI = [0.83, 0.88]) and identification with cause supporters (e.g. "I identify with other supporters of the prochoice position"; adapted from Thomas et al., 2016) using three items and on a seven-point scale ($\alpha = 0.83$, 95% CI = [0.81, 0.86]). The order of presentation of the above measures was randomized. Participants then completed a measure of the mediating mechanism explored in SOM-III: people's essentialist beliefs about a cause ($\alpha = 0.92$, 95% CI = [0.90, 0.93]); Bastian and Haslam, 2006). Finally, they provided demographic information before exiting the survey. No mention was made of the second session of the study. Means, standard deviations, and inter-variable correlations are reported in Table 2.

4.4. Time 2 (T2)

4.4.1. Participants

Approximately two weeks later, the second session of the study, titled "Comment Moderation Task", was posted. Only participants who completed the T1 survey could view the posting, but they were not aware of this, and the study posting did not describe the eligibility criterion or its connection to the first part of the study. Under these circumstances, it is highly likely that participants perceived no connection between the first and second session of the study. A total of 542 participants completed the second session of the study. Two

Table 2
Means, standard deviations, and correlations of measures in Study 2 (N = 540).

Variable	*M*	*SD*	1	2	3	4	5	6	7	8
1. Fusion with cause	4.48	1.44								
2. Attitude extremity	8.14	1.22	0.39**							
3 Attitude centrality	7.27	1.92	0.51**	0.52**						
4. Attitude certainty	8.17	1.22	0.43**	0.69**	0.57**					
5. Attitude importance	7.09	1.96	0.64**	0.58**	0.66**	0.60**				
6. Moral conviction	3.76	1.10	0.49**	0.43**	0.72**	0.47**	0.55**			
7. Identification with cause supporters	6.25	1.03	0.52**	0.72**	0.48**	0.58**	0.58**	0.47**		
8. Rate of censoring congruent comments	0.21	0.16	−0.03	−0.16**	−0.12**	−0.15**	−0.08	−0.11*	−0.17**	
9. Rate of censoring incongruent comments	0.32	0.23	0.11**	0.06	0.02	0.08	0.09*	0.03	0.01	0.53**

Note. The censoring rates, ranging from 0 to 1, refer to the proportion of comments of each type (congruent, incongruent, or irrelevant) that participants censored. Fusion's effect on selective censoring is the difference between fusion's association with the censoring rates of congruent and incongruent comments. This effect was not moderated by position on abortion rights. * indicates $p < .05$. ** indicates $p < .01$.

participants who completed less than 50% of the task were excluded,[4] leaving us with a final sample of 540 participants (48.70% female; 68.83% White; M_{age} = 33.53; SD_{age} = 12.30; 440 pro-choice and 100 pro-life participants). A sensitivity analysis using Monte Carlo simulations revealed that our sample had 99.8% power to detect the fusion effect on selective censoring reported in Study 1. There were no differences in fusion ($t(743) = 1.19$, $p = .23$, $d = 0.10$) between those who did vs. did not complete the second session of the study.

4.4.2. Comment moderation procedure

Participants read about an online forum for discussions on current affairs. They learned that the forum's administrators had received complaints about inappropriate posts by some users and that their task was to help the administrators identify inappropriate posts and block people who repeatedly posted such content. Participants also learned that the comments and users flagged by them would be removed from the forum by its moderators. Because the study was posted on Prolific using a lab account that had previously been used to post other research studies, participants may have easily linked the task to our university and thus may have felt skeptical about our claims that they were evaluating comments from an actual discussion forum and that their evaluations would have real-world consequences. To address this, the study description said that users of the forum were college students and that the forum was owned and run by our university.

Participants then read a series of 28 comments on the abortion rights issue. The comments were designed to look like screenshots of posts from an actual online discussion forum (see Fig. 3 for an example). As shown in the figure, a user icon and handle were displayed next to each comment. The comments that participants read were systematically varied on two factors: Each comment was either pro-choice or pro-life and either offensive or inoffensive. Of the 28 comments, 14 were pro-choice and 14 were pro-life; 14 were pre-determined via the pilot study to be offensive and 14 were inoffensive. In sum, there were four types of comments ($N = 7$ for each type): Offensive-Prochoice, Inoffensive-Prochoice, Offensive-Prolife, and Inoffensive-Prolife. The pro-choice comments were all posted by a single user, and the pro-life comments were all posted by another user. For each comment, participants could recommend deletion or retention, which was our primary measure of censoring. After evaluating all comments, participants were also asked whether the two users whose comments they read should be banned from the blog ("Ban this user from the blog" or "Do not ban this user from the blog"). Finally, participants were asked about the extent to which they doubted the veracity of our claims on a five-point scale (1 = *Not at all*; 5 = *A great deal*), and the mean rating ($M = 2.56$, $SD = 0.98$) was lower than the mid-point of the scale (i.e., 3 = A moderate amount; $t(533) = -10.282$, $p < .001$, $d = -0.45$), suggesting that a considerable proportion of participants believed that they were helping the moderators of a real blog.

For each participant, we calculated censoring rates corresponding to the proportion of comments congruent with the participant's position on abortion rights and also the proportion of incongruent comments that they flagged. As in Study 1, selective censoring was indicated by a higher censoring rate for incongruent than congruent comments. For the offensiveness-related analyses, we also computed censoring rates for each of the four types of comments (Offensive-Congruent, Offensive-Incongruent, Inoffensive-Congruent, and Inoffensive-Incongruent) to determine whether participants selectively censored incongruent comments among both offensive and inoffensive comments. Overall, participants censored offensive comments ($M = 0.47$, $SD = 0.29$) more than inoffensive comments ($M = 0.06$, $SD = 0.13$; $t(559) = 35$, $p < .001$, $d = 1.79$) indicating that the offensiveness manipulation was successful. The censoring rates for offensive and inoffensive comments were correlated [$r(538) = 0.27, p < .001$], indicating that there are relatively stable individual differences in participants' censoring rates.

5. Study 2 results

5.1. Did people selectively censor comments incongruent with their cause and the comments' authors?

Although not pre-registered, we tested whether people generally selectively censored incongruent comments more than congruent comments We compared the censoring rates for incongruent vs. congruent comments via a paired *t*-test. Replicating Study 1 findings, people censored 32.40% ($SD = 22.88$) of the incongruent comments but only 20.64% ($SD = 16.18\%$) of the congruent comments, $t(539) = 13.84$, $p < .001$, $d = 0.58$.

We also conducted exploratory analysis testing whether people were disproportionately willing to ban the author of the incongruent comments relative to the author of the congruent comments. We used a McNemar's Chi-squared test to account for the within-subjects nature of the data and found a significant effect ($\chi^2(1) = 9.24$, $p = .002$) such that 21.31% of participants opted to ban the user who posted incongruent comments as opposed to just 15.41% who banned the user posting congruent comments.

5.2. Did identity fusion amplify the selectively censoring of incongruent comments and their authors?

5.2.1. Selectively censoring of incongruent comments

To test our pre-registered hypothesis that strongly fused individuals would be especially likely to selectively censor incongruent comments, we tested a SEM model similar to Study 1 (see Fig. 4) with residual

[4] In Studies 2 and 3, we excluded participants who responded to fewer than 50% of the comments because their censoring rates are likely to be inaccurate estimates. Note that this exclusion criterion was not pre-registered. In both studies, including these participants did not alter our findings.

PostalExplorer

4 days ago

Although it is not explicitly stated in the constitution, I strongly believe in a women's right to an abortion. Outlawing abortion restricts women's freedom.

Fig. 3. Example of an inoffensive pro-choice comment used in the comment moderation task (Study 2).

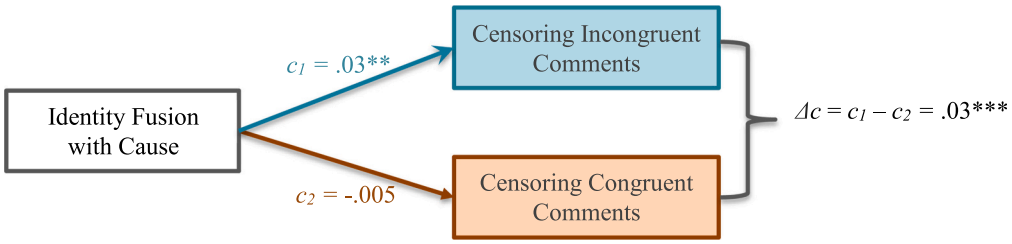

Fig. 4. Structural Equations Model depicting the effect of identity fusion on selective censoring of incongruent vs. congruent comments (Study 2). The c_1 and c_2 paths represent the effects of fusion on censoring incongruent and congruent comments respectively. The path coefficients in the figure are unstandardized. The significant difference between the two paths (Δc) indicates that fusion is associated with selectively censoring incongruent comments. ** indicates $p < .01$. *** indicates $p < .001$.

Table 3
Path coefficients (c_1 and c_2) and Chi-sq values (χ^2) of SEM models and coefficients from regression models testing the effects of each identity-related measure on selective censoring (Study 2). Note that each model included only one predictor.

Predictor in model	Semantic equation modeling (SEM)			Selective Censoring difference index
	Censoring incongruent comments (c_1)	Censoring congruent comments (c_2)	Selective censoring ($\Delta c = c_1$-c_2) χ^2	(b)
Model 1: Fusion with cause	0.03**	−0.005	13.14***	0.03***
Model 2: Attitude importance	0.02*	−0.01†	15.09***	0.03***
Model 3: Attitude certainty	0.02†	−0.025***	25.25***	0.04***
Model 4: Attitude centrality	0.004	−0.02**	7.35**	0.02**
Model 5: Attitude extremity	0.01	−0.025***	20.095***	0.04***
Model 6: Identification with cause supporters	0.002	−0.03***	11.595***	0.03***
Model 7: Moral conviction	0.007	−0.02*	8.68**	0.03**

Note. In each model, the predictor was standardized, but the censoring rates were not. The censoring rates ranged from 0 to 1. The path coefficients reported are unstandardized. † indicates $p < .1$. * indicates $p < .05$. ** indicates $p < .01$. *** indicates $p < .001$.

covariances between the censoring rates. Alternate analyses treating the difference between censoring rates of incongruent and congruent comments as the selective censoring index did not alter our conclusions (see the last column in Table 3 in the article and SOM-II). As in Study 1, we standardized the continuous predictors in all the SEM analyses, and we report unstandardized regression coefficients. Fusion positively predicted censoring incongruent comments (c_1 path; $b = 0.03$, 95% CI = [0.01, 0.045], $p = .008$) but not censoring congruent comments (c_2 path; $b = -0.005$, 95% CI = [−0.02, 0.01], $p = .50$). Replicating Study 1's main finding, the difference between the fusion effects on censoring incongruent vs. congruent comments was statistically significant, ($\Delta c = c_1 - c_2$; $\chi^2(1) = 13.14$, $p < .001$), which is evidence that fusion is associated with selective censoring. To illustrate, participants who were strongly fused (+ 1 *SD*) censored 36.36% of the incongruent comments they read but only 18.65% of the congruent comments. Weakly fused participants censored 29.49% of the incongruent comments and 21.26% of the congruent comments, indicating a weaker selective censoring tendency. Fusion's effect on selective censoring remained significant when we controlled for whether participants were pro-choice or pro-life ($\chi^2(1) = 13.50$, $p < .001$), and the effect was not moderated by position on abortion rights ($\chi^2(1) = 0.04$, $p = .85$).

Our pre-registered mediational analyses (see SOM-III) suggest that essentialistic beliefs regarding people's stance on abortion rights might be at least one mediating mechanism explaining the fusion effect on selective censoring. In our pre-registration, we also proposed to test the fusion effect controlling for other identity-related measures. We accordingly report a model in which the predictive ability of all the identity-related measures are compared (see SOM-V). Nevertheless, because the measured variables are all strongly related both conceptually and empirically (see Table 2), after establishing that multicollinearity was not a problem, we examined whether each of these variables independently predicts selective censoring.

5.2.2. Selective censoring of the authors of incongruent comments

The foregoing analyses revealed that identity fusion with a cause is associated with a tendency to disproportionately censor online content that is incongruent with the cause. To test the pre-registered hypothesis that strongly fused individuals would also display a censoring bias against the authors of incongruent content, we examined a SEM model with two dependent variables corresponding to the binary indicators of whether the participant decided to ban the authors of incongruent, and congruent comments. Fusion was not significantly associated with banning the author of the incongruent comments (OR = 1.17, 95% CI = [0.95, 1.45], $p = .14$) or congruent comments (OR = 0.99, 95% CI = [0.78, 1.25], $p = .90$). The difference between the two paths, computed as two times the negative loglikelihood of the difference between the two paths, was not significant ($\chi^2(1) = 1.18$, $p = .28$), indicating that fusion was not associated with selectively censoring authors of incongruent comments. However, given that the non-

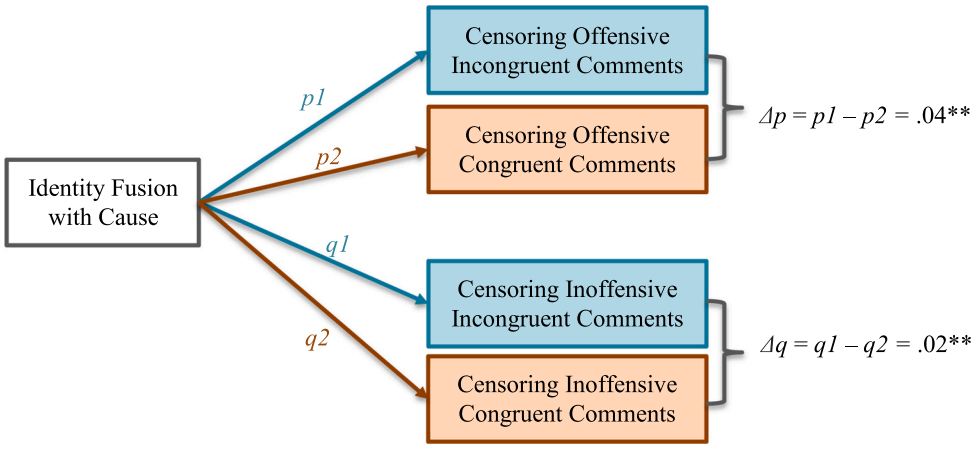

Fig. 5. Structural Equations Model examining the effect of identity fusion on selective censoring of incongruent vs. congruent comments among offensive and inoffensive comments (Study 2). Δ*p* and Δ*q* represent fusion's effects on selective censoring among offensive comments and inoffensive comments, respectively. The significant effects indicate that strongly fused people selectively censored incongruent comments whether the comments were offensive or inoffensive. See SOM-IV for path coefficients. * indicates $p < .05$. ** indicates $p < .01$.

significant coefficients of the two paths were in the predicted direction, it is possible that there exists a small effect that our sample was not sufficiently powered to detect.

5.2.3. Did offensiveness moderate the effect of fusion on selectively censoring?

To verify Study 1's exploratory finding and our pre-registered hypothesis that the offensiveness of comments would not moderate the effect of fusion on selective censoring, we modeled the paths from fusion to participants' censoring rates for four types of comments: Offensive-Congruent, Offensive-Incongruent, Inoffensive-Congruent, and Inoffensive-Incongruent (see Fig. 5).

Among offensive comments, fusion was associated with selectively censoring incongruent comments over congruent comments ($\Delta p = p1 - p2$; $b = 0.04$, 95% CI = [0.02, 0.06], $p = .001$). Similarly, among inoffensive comments, strongly fused individuals selectively censored incongruent comments ($\Delta q = q1 - q2$; $b = 0.02$, 95% CI = [0.005, 0.04], $p = .008$). (The four path coefficients are reported in SOM-IV). The two significant selective censoring effects suggest that strongly fused people's selective intolerance for incongruent comments was observable among both offensive and inoffensive comments. Comparing two selective censoring effects for offensive vs. inoffensive comments ($\Delta p - \Delta q$) revealed a marginally significant difference ($\chi^2(1) = 3.34$, $p = .07$), suggesting that fusion's effect on selective censoring may have been larger for offensive than inoffensive comments. What is striking however is that as in Study 1, strongly fused people selectively censored incongruent comments even when the comments were inoffensive.

5.3. Did fusion's effect on selective censoring of incongruent comments generalize to other identity-related measures?

Thus far, we focused on the effects of identity fusion. Nevertheless, we conducted exploratory analyses testing the possibility that selective censoring of incongruent comments results from a constellation of identity-related processes. To test this possibility, we assessed the effects of attitude strength (attitude extremity, attitude centrality, attitude certainty, and attitude importance), moral conviction, and identification with supporters, which all index different aspects of people's alignment with a cause. Using the same approach as in the fusion analysis, we sequentially tested the relation of each of the seven predictors to selective censoring. Table 3 reports each model's path coefficients from the tested variable to censoring incongruent comments (c_1) and to censoring congruent comments (c_2). Table 3 also reports the chi-square difference between the two paths ($c_1 - c_2$) indicating the extent to which the tested variable is associated with selectively censoring incongruent comments. The last column presents linear regression coefficients from alternate analyses testing the effect of each identity-related measure on the difference in participants' censoring rates for incongruent vs. congruent comments.

As indicated by the significant chi-square differences (Δc) and the significant regression coefficients (b) in Table 3, each of the constructs produced selective censoring similar to the fusion effects, which is preliminary evidence that broader identity-related processes motivate selective censoring.

Interestingly, most of the predictors (attitude certainty, attitude centrality, attitude extremity, identification with cause supporters, and moral conviction) were negatively associated with censoring congruent comments (see c_2 coefficients in Table 3), indicating that they produce a tendency to be lenient toward congruent comments. On the contrary, fusion and attitude importance were not correlated with censoring congruent comments; instead, they were positively associated with censoring incongruent comments (see c_1 coefficients in Table 3), implying that these constructs were associated with an intolerance for incongruent comments. We speculate that a preference for congruent content and an intolerance against incongruent content reflect two independent mechanisms leading to selective censorship of incongruent comments.

5.4. Did selective censoring of incongruent comments depend on people's ideologies?

We tested another SEM model (not pre-registered) similar to the fusion analysis to assess the effect of people's stance on abortion rights (pro-choice vs. pro-life). Unlike Study 1, pro-choice participants selectively censored incongruent comments as much as pro-life participants ($\chi^2(1) = 2.38$, $p = .12$), which may be due to higher threat levels among pro-choice participants following the, 2018 nomination Justice Kavanaugh to the Supreme Court. That is, owing to the conservative shift in the makeup of the Supreme Court in, 2018, pro-choice participants in Study 2 may have generally faced higher threat relative to Study 1, which could have increased their tendency to selectively censor pro-life comments. There was also no difference in fusion levels among pro-choice and pro-life participants ($t(537) = 0.59$, $p = .56$, $d = 0.07$).

6. Study 2 discussion

Study 2 replicated Study 1's main findings that people censor online content that is incongruent with their own political views and that strongly fused individuals are especially likely to selectively censor incongruent content. Strongly fused people's tendency to selectively censor incongruent comments was robust for both offensive and inoffensive comments. Contrary to Study 1, we did not find evidence that pro-life participants selectively censored more than pro-choice participants, which we believe could be due to the socio-political environment during Study 2 data collection.

In addition to replicating Study 1 effects, Study 2 also examined people's willingness to ban the authors of incongruent vs. congruent comments from the forum. We found that cause supporters selectively banned the author who consistently posted cause-incongruent content. Contrary to our hypothesis, this effect was not amplified by fusion. This may have been because banning authors is a relatively extreme action that participants in our samples generally did not endorse. Conceivably, there is a small association of fusion with selective censoring of authors that our sample was underpowered to detect.

Finally, the study found that the selective censoring effect extends to an array of identity-related measures in the literature. The findings also indicate that there may be different paths to selective censorship of opposing content: Whereas fusion and attitude importance were associated with an increased tendency to censor incongruent comments, the other identity-related predictors were associated with a weaker tendency to censor congruent comments.

In short, the results of Study 2 replicated the selective censoring effect that emerged in Study 1. A potential limitation of these studies, however, is that both focused on an issue rooted in religious values, abortion rights. To address this, Study 3 focused on gun rights. The gun-rights issue was particularly relevant in the time that the study was conducted because gun sales peaked during the COVID-19 crisis (Collins and Yaffe-Bellany, 2020).

7. Study 3

The method used in Study 3 resembled those used in previous studies except that we used a more controlled manipulation of comment offensiveness that kept the content of the comments constant. Whereas in Study 2 comments were categorized as offensive or inoffensive based on coders' ratings, in Study 3, for each inoffensive comment, we generated an offensive version by adding offensive phrases. In this way, the content of inoffensive and comments was identical except for offensive language. Finally, as in Study 2, we assessed whether the selective censoring effect of fusion generalized to other identity-related measures such as indices of attitude strength, moral conviction, and identification with cause supporters.

8. Study 3 Method

8.1. Power analysis

As mentioned in our pre-registration (see https://osf.io/x3w7h/?view_only=a25d722f3a03405e9e4f074a622b10b4), an a priori power analysis conducted using Monte Carlo simulations indicated that a sample of 325 participants was required to detect the selective censoring effect detected in Study 2 with an alpha of 0.05 and 80% power. Given the longitudinal nature of the study, we estimated that approximately 30% of the sample would either drop out between T1 and T2 or fail attention checks, and so we decided to recruit 460 participants at T1.

8.2. Time 1 (T1)

8.2.1. Participants

A sample of 466 participants (49.6% female; 67.0% White; M_{age} = 31.18; SD_{age} = 11.14) from Prolific Academic completed the first part of the study in May 2020. Participants' views on gun rights were measured in the T1 survey (370 pro-gun-control and 96 pro-gun-rights participants).

8.2.2. Identity measures

Participants completed the identity fusion scale for their position on gun rights (either pro-gun or anti-gun) on a seven-point scale (α = 0.93). Using the measures used in Study 2, we measured four facets of attitude strength – attitude extremity, attitude centrality, attitude certainty and attitude importance, moral conviction, and identification with cause supporters (α = 0.86). The order of presentation of the above constructs was randomized. Means, standard deviations, and inter-variable correlations are reported in Table 5. Finally, they provided demographic information.

8.3. Time 2 (T2)

8.3.1. Participants

Two weeks after completing the T1 survey, participants were able to complete a "Comment Moderation Task". A total of 373 participants completed the task. Two participants who completed less than 50% of the task were excluded, leaving us with a final sample of 371 participants (52.85% female; 66.85% White; M_{age} = 31.45; SD_{age} = 11.61; 297 pro-gun-control and 74 pro-gun-rights participants). A sensitivity analysis revealed that our sample had 85% power to detect the fusion effect on selective censoring reported in Study 2. We found a difference in fusion levels between people who did vs. did not complete the T2 session such that individuals who completed T2 were more fused with the cause ($t(462) = 2.01$, $p = .05$, $d = -0.23$).

8.3.2. Comment moderation procedure

As in the previous studies, we asked participants to help moderators of a college-run discussion forum identify inappropriate posts for removal. We gathered 14 pro-gun-rights comments and 14 pro-gun-control comments from the internet, resulting in 28 comments. We created offensive and inoffensive versions of each comment by including or excluding offensive phrases. Participants read either the offensive or inoffensive version of each of the 28 comments. Overall, participants read four types of comments (N = 7 for each type): Offensive-Pro-gun-rights, Inoffensive-Pro-gun-rights, Offensive-Pro-gun-control, and Inoffensive-Pro-gun-control (See Table 4 for example comments). As in Study 2, each comment was accompanied by a user icon and timestamp like in real online forums. The pro-gun-rights comments were all posted by a single user, and the pro-gun-control comments were all posted by another user. As in the previous studies, for each comment, participants recommended deletion or retention. After evaluating all comments, participants were also asked whether the two users whose comments they read should be banned from the blog ("Ban this user from the blog" or "Do not ban this user from the blog"). Finally, participants rated how much they doubted that the forum was not real on a five-point scale (1 = not at all, 5 = a great deal). The mean rating (M = 2.65, SD = 0.99) was lower than the mid-point of the scale (i.e., 3 = A moderate amount; $t(366) = -6.77$, $p < .001$, $d = -0.35$), suggesting that participants generally did not doubt the veracity of the paradigm.

For each participant, we calculated censoring rates corresponding to comments congruent and incongruent with their own position on guns. For the offensiveness-related analyses, we also computed censoring rates for each of the four types of comments (Offensive-Congruent, Offensive-Incongruent, Inoffensive-Congruent, and Inoffensive-Incongruent). Overall, participants censored offensive comments (M = 0.58, SD = 0.28) more than inoffensive comments (M = 0.07, SD = 0.12; $t(370) = 33.98$, $p < .001$, $d = 2.27$) indicating that the offensiveness manipulation was successful. The censoring rates for offensive and inoffensive comments were correlated albeit more weakly than in Study 1 ($r(369) = 0.17$, $p < .001$).

9. Study 3 results

9.1. Did people selectively censor comments incongruent with their cause and the comments' authors?

We tested the pre-registered hypothesis that people would selectively censor incongruent comments more than congruent comments. We conducted a paired t-test comparing the censoring rates for incongruent vs. congruent comments. Replicating findings from the first two

Table 4
Sample comments rated by participants in Study 3. The study included 28 comments (14 pro-gun-rights and 14 pro-gun-control), each of which had an offensive and an inoffensive version. Participants rated either the offensive or inoffensive version of each of the 28 comments. The comments were presented in the format illustrated in Fig. 3 and in random order.

	Sample comments rated by Participant 1	Sample comments rated by Participant 2
Pro-gun-rights	*PostalExplorer*: We must defend the right to keep and bear arms through communication and coordinated action, retarded dumbasses like you just don't get it. [**offensive**]	*PostalExplorer*: We must defend the inherent right to keep and bear arms through communication and coordinated action. [**inoffensive**]
	PostalExplorer: Everyone should be pro gun. Pro gun = pro freedom. Pro gun = anti tyranny. [**inoffensive**]	*PostalExplorer*: You're must be an unfixable dumbfuck if you don't get this: Pro gun = pro freedom. Pro gun = anti tyranny. [**offensive**]
Pro-gun-control	*Emerald-3*: Why aren't guns and, oh yeah, assault rifles banned? Why aren't you banned? It is unbelievable that this has been allowed to continue. I am mortified that you exist. Enough is enough! #guncontrol #fuckguns [**offensive**]	*Emerald-3*: Why aren't guns and specifically assault rifles banned? It is unbelievable that this has been allowed to continue. Enough is enough! #guncontrol #nomoreguns [**inoffensive**]
	Emerald-3: I don't care about Thoughts and Prayers. It's just a phrase that people use instead of "Thoughts and Actions". [**inoffensive**]	*Emerald-3*: I Don't Give a Fuck About Your Thoughts and Prayers. It's just a shitty, waste of words that people use instead of "Thoughts and Actions". [**offensive**]

studies, people censored more incongruent comments (M = 36.97%; SD = 19.64) than congruent comments (M = 27.88%; SD = 17.62), t (370) = 10.02, p < .001, d = 0.49.

We also conducted a pre-registered analysis testing whether people were disproportionately willing to ban the author of the incongruent comments relative to the author of the congruent comments. Contrary to our hypothesis and the results of Study 1, we did not find a significant difference ($\chi^2(1) = 1.92$, $p = .17$). Nevertheless, the means trended in the expected direction. That is, 32.69% of participants banned the user who posted incongruent comments as opposed to just 29.51% who banned the user posting congruent comments.

9.2. Did identity fusion amplify the selectively censoring of incongruent comments?

To test our pre-registered hypothesis that strongly fused individuals would be especially likely to selectively censor incongruent comments, we tested a SEM model (see Fig. 6) with residual covariances between the censoring rates. (Alternate analyses treating the difference between censoring rates of incongruent and congruent comments as the selective censoring index, reported in Table 6 below and in SOM-II, result in the same findings). As in Studies 1 and 2, we standardized the predictors in all the SEM analyses, and we report unstandardized regression coefficients. Fusion positively (but not significantly) predicted censoring incongruent comments (c_1 path; b = 0.02, 95% CI = [−0.004, 0.04], p = .12) but not censoring congruent comments (c_2 path; b = −0.006, 95% CI = [−0.02, 0.01], p = .49). The difference between the fusion effects on censoring incongruent vs. congruent comments was significant, ($\Delta c = c_1 - c_2$; $\chi^2(1) = 6.01$, $p = .01$), which is evidence that fusion is associated with selective censoring. To illustrate, participants who were strongly fused (+ 1 SD) censored 41.47% of the incongruent comments they read but only 28.56% of the congruent comments. Weakly fused participants censored 35.92% of the incongruent comments and 29.52% of the congruent comments, indicating weaker selective censoring. The effect of fusion on selective censoring remained significant when we controlled for whether participants favored pro-gun-rights or pro-gun-control ($\chi^2(1) = 9.24$, $p = .002$), and the effect was not moderated by position on gun rights ($\chi^2(1) = 0.05$, $p = .83$).

9.2.1. Did offensiveness moderate the effect of fusion on selectively censoring?

As in the previous studies and consistent with the pre-registration, we modeled the paths from fusion to participants' censoring rates for four types of comments: Offensive-Congruent, Offensive-Incongruent, Inoffensive-Congruent, and Inoffensive-Incongruent (see Fig. 7). Among inoffensive comments, fusion was associated with selectively censoring incongruent comments over congruent comments ($\Delta q = q1 - q2$; b = 0.03, 95% CI = [0.009, 0.04], p = .003). Among offensive comments, the effect was in the predicted direction but not significant ($\Delta p = p1 - p2$; $b = 0.02$, 95% CI = [−0.007, 0.04], p = .16). (The four path coefficients are reported in SOM-IV). Comparing two selective censoring effects for offensive vs. inoffensive comments ($\Delta p - \Delta q$) revealed no difference ($\chi^2(1) = 0.39$, $p = .53$).

Table 5
Means, standard deviations, and correlations with confidence intervals in Study 3 (N = 371).

Variable	*M*	*SD*	1	2	3	4	5	6	7	8
1. Fusion with cause	3.45	1.43								
2. Attitude extremity	7.62	1.36	0.31**							
3 Attitude centrality	6.75	1.84	0.49**	0.53**						
4. Attitude certainty	7.58	1.46	0.38**	0.69**	0.55**					
5. Attitude importance	6.55	1.79	0.61**	0.55**	0.73**	0.59**				
6. Moral conviction	3.37	1.03	0.49**	0.49**	0.68**	0.54**	0.56**			
7. Identification with cause supporters	5.65	1.06	0.50**	0.63**	0.56**	0.63**	0.61**	0.57**		
8. Rate of censoring congruent comments	0.28	0.18	−0.04	−0.01	−0.05	−0.01	−0.04	−0.06	−0.04	
9. Rate of censoring incongruent comments	0.37	0.20	0.08	0.11*	0.10*	0.13**	0.13*	0.07	0.07	0.56**

Note. The censoring rates, ranging from 0 to 1, refer to the proportion of comments of each type (congruent and incongruent) that participants censored. Fusion's effect on selective censoring is the difference between fusion's association with the censoring rates of congruent and incongruent comments. This effect was not moderated by position on gun rights. * indicates p < .05. ** indicates p < .01.

Table 6
Path coefficients (c_1 and c_2) and Chi-sq values (χ^2) of SEM models and coefficients from regression models testing the effects of each identity-related measure on selective censoring (Study 3). Note that each model included only one predictor.

Predictor in model	Semantic equation modeling (SEM)			Selective Censoring difference index (b)
	Censoring incongruent comments (c_1)	Censoring congruent comments (c_2)	Selective censoring ($\Delta c = c_1$-c_2) χ^2	
Model 1: Fusion with cause	0.02	−0.006	6.01*	0.02*
Model 2: Attitude importance	0.03*	−0.01	13.45***	0.03***
Model 3: Attitude certainty	0.03**	−0.002	9.86**	0.03**
Model 4: Attitude centrality	0.02*	−0.01	11.26***	0.03***
Model 5: Attitude extremity	0.02*	−0.002	7.01**	0.02**
Model 6: Identification with cause supporters	0.02	−0.007	5.51*	0.02*
Model 7: Moral conviction	0.01	−0.01	7.33**	0.03**

Note. In each model, the predictor was standardized, but the censoring rates were not. The censoring rates ranged from 0 to 1. The path coefficients reported are unstandardized. * indicates $p < .05$. ** indicates $p < .01$. *** indicates $p < .001$.

9.3. Did fusion's effect on selective censoring of incongruent comments generalize to other identity-related measures?

We then tested our pre-registered hypothesis that fusion's effect on selective censoring would extend to seven identity-related measures. Using models similar to the fusion analysis, we tested the effect of each predictor on selective censoring. Table 6 reports each model's path coefficients from the tested variable to censoring incongruent (c_1) and congruent (c_2) comments, and the chi-square difference between the two paths ($c_1 - c_2$) indicating the extent to which the tested variable is associated with selective censoring. The last column in Table 6 presents linear regression coefficients from alternate models testing the effect of each identity-related measures on the difference between participants' censoring rates for incongruent and congruent comments. The significant chi-square differences (Δc) and regression coefficients (b) indicate that the selective censoring effect generalized to each of the seven identity-related measures. In contrast to Study 2, the selective censoring effect was largely driven by positive associations between the identity-related measures and censoring incongruent comments.

9.4. Did selective censoring of incongruent comments depend on people's ideologies?

We tested another exploratory SEM model to assess the effect of people's stance on gun rights (pro-gun-rights vs. pro-gun-control). Gun-control supporters selectively censored incongruent comments more than gun-rights supporters ($\chi^2(1) = 17.09$, $p < .001$) even though pro-gun- rights supporters tended to be more strongly fused than pro-gun- control supporters ($t(367) = 2.18$, $p = .03$, $d = 0.28$). Study 3 was conducted during a period that saw increased gun sales (Collins and Yaffe-Bellany, 2020), which should have increased the threat perceived by gun-control supporters, increasing their tendency to selectively censor opposition.\.

10. Study 3 discussion

Study 3 demonstrated that the selective censoring effect extends to issues beyond religiously tinged issues such as abortion rights. Specifically, people selectively censored comments that opposed their views on the gun rights debate, and this effect was amplified among people who were strongly fused with their cause. As in Studies 1 and 2, people selectively censored incongruent comments even when they were inoffensive. Contrary to Study 2, we did not find a significant selective censoring effect on offensive comments, but it could be that our study was underpowered to detect this effect. Further, gun-control proponents selectively censored more than gun-rights proponents, which when taken together with Studies 1 and 2, suggests that people's willingness to selectively censor may depend on the cause at hand (pro-choice or pro-gun-control) and the political context (e.g., level of threat faced by the cause) rather than political ideology (left or right).

Study 3 also replicated the Study 2 finding that selective censoring extends to a range of identity related constructs including attitude strength, identification with supporters, and moral conviction. Nevertheless, we did not find similar results across Studies 2 and 3 regarding the degree to which each identity-related process produced a lenience toward congruent content or an intolerance of incongruent content. Future research will need to disentangle the links between identity related processes and selective censoring.

10.1. General discussion

The current research provides an initial glimpse into how people censor political opponents when moderating online content. Specifically, in three studies, participants who were asked to moderate an online forum deleted approximately 5–12% more identity-incongruent, relative to identity-congruent, comments from putative online forums. Moreover, we found weak evidence that participants were about 3–5% points more likely to ban authors of incongruent as compared to congruent comments. These findings transcend past research on selective exposure and avoidance (Bakshy et al., 2015; Garrett, 2009a; van der Linden, 2017) because censorship is a particularly extreme action that affects not just one's own online environment but also the environments of other people. Furthermore, unlike traditional censorship enforced only by the state (Bonsaver, 2007; Fishburn, 2008), the decentralized nature of this new form of censorship implemented by independent users could make it easy to overlook and thus potentially

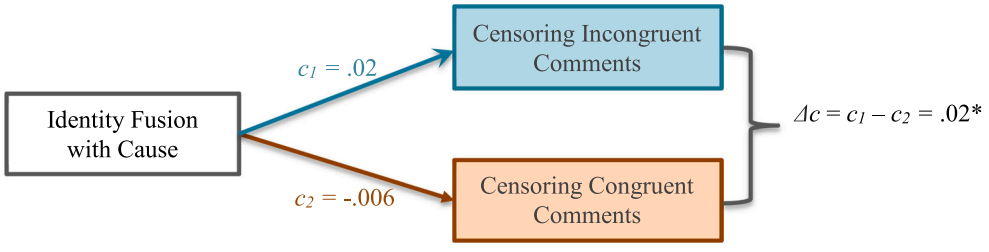

Fig. 6. Structural Equations Model depicting the effect of identity fusion on selective censoring of incongruent vs. congruent comments (Study 3). The c_1 and c_2 paths represent the effects of fusion on censoring incongruent and congruent comments respectively. The significant difference between the two paths (Δc) indicates that fusion is associated with selectively censoring incongruent comments. * indicates $p < .05$.

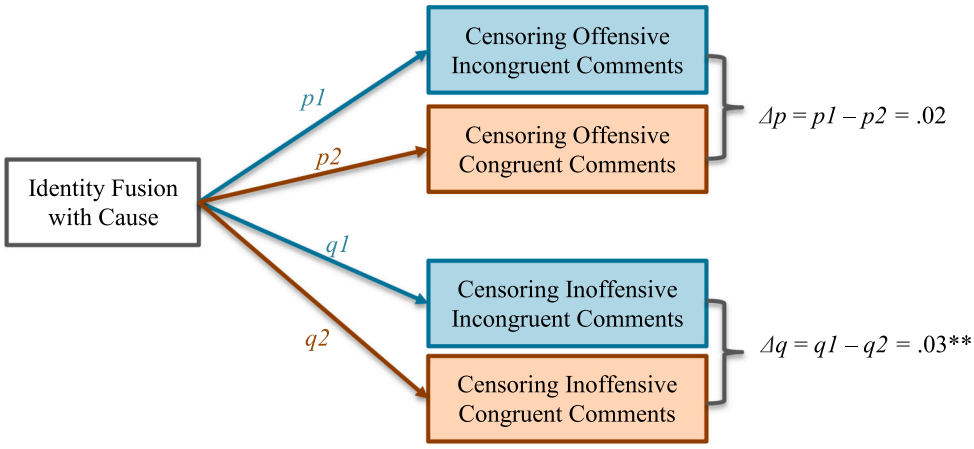

Fig. 7. Structural Equations Model examining the effect of identity fusion on selective censoring of incongruent vs. congruent comments among offensive and inoffensive comments (Study 3). Δ*p* and Δ*q* represent fusion's effects on selective censoring among offensive comments and inoffensive comments, respectively. The difference between them was not significant, which indicates that comment offensiveness did not moderate fusion's effect on selective censoring. See SOM-IV for path coefficients. ** indicates $p < .01$.

more insidious.

Our evidence that people censor the social media posts of political opponents is consistent with recent evidence that the salutary impact of intergroup contact on intergroup harmony (Paluck et al., 2018) may not extend to online interactions (Bail et al., 2018). We also show, however, that selective censorship of opponents' comments was amplified among people whose cause-related views were firmly rooted in their identities. Strongly fused participants deleted approximately 13–18% more identity-incongruent than identity-congruent comments, while weakly fused participants were much less biased (0–9%). Strikingly, strongly fused individuals disproportionately censored opponents' comments even when the comments conveyed opposing views in an inoffensive and courteous manner. The identity-driven effect on selective censoring generalized to six other identity-related measures including indices of attitude strength, moral conviction, and identification with cause supporters. The converging results across the various predictors suggest that selective censoring results from a combination of several identity-related processes.

Future research might work toward developing a theoretical model of selective censoring that elaborates the relationships between various identity-related processes. Such work might also investigate the two possible mechanisms underlying selective censoring: lenience toward congruent content versus intolerance of incongruent content. Future researchers might also follow up on our evidence that strongly fused participants were especially apt to censor opponents' *comments* but not their opponents *themselves*. Also, perhaps people ban individuals based on their most offensive comment rather than based on evaluating multiple comments. Further, whereas we focused on identity-related processes, future research might consider other processes such as expectations regarding the content online subscribers of a given forum prefer (Haselmayer et al., 2017) that may also contribute to moderators' selective censoring.

The censorship effects described here could have considerable impact on online forums and communities that millions of people follow. Studies of moderators have noted that a small number of them govern very large online communities and that they hold enormous power over their communities (Frith, 2014; Matias, 2016b). Still, past work on moderators has largely focused on how people become moderators (Shaw and Hill, 2014), and the nature of their roles (Berge and Collins, 2000; Colladon and Vagaggini, 2017; Frith, 2014) and struggles (Matias, 2016a). Although some case studies have examined abuse of power by moderators (Yang, 2019), including anecdotal evidence of politically motivated censorship (Wright, 2006), the current research is the first systematic investigation of censoring among people who moderate online communities. This investigation is consequential because selective censoring that favors the viewpoints of a small number of moderators could produce huge biases in the content that millions see. Indeed, censoring by powerful moderators can give onlookers who are not aware that censoring has occurred a false sense of the views of the people in an online community and who belongs there.

Still, our findings may generalize beyond the groups of people who serve as moderators of large online communities or forums. The millions of people who own blogs, YouTube channels, and social media pages, can moderate others' comments on the platforms they control. Even regular social media users can moderate others' comments on their own posts. Of course, in our studies, participants were explicitly given the goal of deleting inappropriate comments. Because most regular social media users may not experience a strong deletion-focused goal, they may censor less than moderators do. Nevertheless, the collective impact of each of these individuals' censoring could produce substantial consequences.

We believe censorship is a potentially overlooked factor in the heightened political polarization our culture is witnessing. This could have important ramifications. For example, selective censoring could lead to a lack of exposure to different viewpoints, creating echo chambers and causing people to develop increasingly extreme opinions (Price et al., 2006) and to overestimate the prevalence of their own viewpoints (Ross et al., 1977). In addition, opponents of causes may witness the increased extremism of inhabitants of the echo chamber and respond in kind by adopting extreme opposing views of their own (Bail et al., 2018). These processes may reinforce themselves, producing more and more polarization over time (Allcott et al., 2020). Censorship could also have implications for the people being censored, who may feel marginalized and become disengaged from the online community or be less likely to share his or her views in the future. Future studies should examine the consequences of selective censoring in online contexts.

11. Conclusion

Contemporary pundits often blame the apparent increase in polarization on “the internet” or “social media.” Researchers have found some basis for such assertions by demonstrating that internet users are indeed selectively exposed to evidence that would lend support to their views. Our findings move beyond this literature by demonstrating that moderators employ censorship to not only bring online content into harmony with their values, but to actively advance their causes and attack opponents of their causes. From this vantage point, those whose political beliefs are rooted in their identities are not passive participants in online polarization; rather, they are agentic actors who actively curate online environments by censoring content that challenges their ideological positions. By providing a window into the psychological processes underlying these processes, our research may open up a broader vista of related processes for systematic study.

Funding

This work was supported by the National Science Foundation [grants BCS-1124382 and BCS1528851 to William B. Swann, Jr.], an Advanced Grant from the European Research Council 694986 to Michael Buhrmester, and grant by Ministerio de Ciencia, Innovación y Universidades RTI2018-093550-B-I00 to Angel Gomez. The funders played no role in the study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Acknowledgments

We thank Elliot Tucker-Drob and Greg Hixon for their help with the data analysis.

Open practices

All study materials and data used in this research have been made publicly available and can be accessed at https://osf.io/4jtwk/?view_only=10627a9892464e5aa90fe92360b846ad. The design, methods, and analysis plan of Studies 2 and 3 were pre-registered, and these can be viewed at https://osf.io/2jvau?view_only=754165d77cbe4e69baf6b11740b1a422 and https://osf.io/x3w7h/?view_only=a25d722f3a03405e9e4f074a622b10b4 respectively.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2020.104031.

References

Adamic, L. A., & Glance, N. (2005, August). The political blogosphere and the 2004 US election: divided they blog. *Proceedings of the 3rd international workshop on Link discovery* (pp. 36–43). ACM.

Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review, 110*(3), 629–676.

Ashokkumar, A., Galaif, M., & Swann, W. B., Jr. (2019). Tribalism can corrupt: Why people denounce or protect immoral group members. *Journal of Experimental Social Psychology, 85*, 103874.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences, 115*(37), 9216–9221.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science, 348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science, 26*(10), 1531–1542.

Bastian, B., & Haslam, N. (2006). Psychological essentialism and stereotype endorsement. *Journal of Experimental Social Psychology, 42*(2), 228–235. https://doi.org/10.1016/j.jesp.2005.03.003.

Berge, Z. L., & Collins, M. P. (2000). Perceptions of e-moderators about their roles and functions in moderating electronic mailing lists. *Distance Education, 21*(1), 81–100.

Binder, A. R., Dalrymple, K. E., Brossard, D., & Scheufele, D. A. (2009). The soul of a polarized democracy: Testing theoretical linkages between talk and attitude extremity during the 2004 presidential election. *Communication Research, 36*(3), 315–340. https://doi.org/10.1177/0093650209333023.

Boninger, D. S., Krosnick, J. A., & Berent, M. K. (1995). Origins of attitude importance: Self-interest, social identification, and value relevance. *Journal of Personality and Social Psychology, 68*(1), 61. https://doi.org/10.1037/0022-3514.68.1.61.

Bonsaver, G. (2007). *Censorship and literature in fascist Italy*. University of Toronto Press.

Boutyline, A., & Willer, R. (2017). The social structure of political echo chambers: Variation in ideological homophily in online networks. *Political Psychology, 38*(3), 551–569.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, 114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114.

Brewer, M. B. (2001). Ingroup identification and intergroup conflict. *Social Identity, Intergroup Conflict, and Conflict Reduction, 3*, 17–41.

Clarkson, J. J., Tormala, Z. L., DeSensi, V. L., & Wheeler, S. C. (2009). Does attitude certainty beget self-certainty? *Journal of Experimental Social Psychology, 45*(2), 436–439. https://doi.org/10.1016/j.jesp.2008.10.004.

Colladon, A. F., & Vagaggini, F. (2017). Robustness and stability of enterprise intranet social networks: The impact of moderators. *Information Processing & Management, 53*(6), 1287–1298.

Collins, K., & Yaffe-Bellany, D. (2020, April 2). About 2 million guns were sold in the US as virus fears spread. *The New York times*. Retrieved from www.nytimes.com.

Conditt, J. (2016, July 28). Moderators banned 2,200 accounts during Donald Trump's AMA. *Engadget*. Retrieved from https://www.engadget.com.

Crawford, J. T., & Pilanski, J. M. (2014). Political intolerance, right and left. *Political Psychology, 35*(6), 841–851.

Fazio, R. H., & Zanna, M. P. (1978). Attitudinal qualities relating to the strength of the attitude-behavior relationship. *Journal of Experimental Social Psychology, 14*(4), 398–408. https://doi.org/10.1016/0022-1031(78)90035-5.

Fishburn, M. (2008). The burning of the books. *Burning Books* (pp. 31–48). London: Palgrave Macmillan.

Fisher, R., Lilie, S., Evans, C., Hollon, G., Sands, M., Depaul, D., ... Hultgren, T. (1999). Political ideologies and support for censorship: Is it a question of whose ox is being gored? *Journal of Applied Social Psychology, 29*(8), 1705–1731.

Fredman, L. A., Bastian, B., & Swann, W. B., Jr. (2017). God or country? Fusion with Judaism predicts desire for retaliation following Palestinian stabbing Intifada. *Social Psychological and Personality Science, 1948550617693059*. https://doi.org/10.1177/1948550617693059.

Frith, J. (2014). Forum moderation as technical communication: The social web and employment opportunities for technical communicators. *Technical Communication, 61*(3), 173–184.

Garimella, V. R. K., & Weber, I. (2017, May). A long-term analysis of polarization on Twitter. *Eleventh international AAAI conference on web and social media*.

Garrett, R. (2008). Selective processes, exposure, perception, memory. In L. L. Kaid, & C. Holtz-Bacha (Vol. Eds.), *Encyclopedia of political communication. Vol. 1. Encyclopedia of political communication* (pp. 741–). Thousand Oaks, CA: SAGE Publications, Inc. https://doi.org/10.4135/9781412953993.n619.

Garrett, R. K. (2009a). Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of Computer-Mediated Communication, 14*(2), 265–285. https://doi.org/10.1111/j.1083-6101.2009.01440.x.

Garrett, R. K. (2009b). Politically motivated reinforcement seeking: Reframing the selective exposure debate. *Journal of Communication, 59*(4), 676–699.

Gómez, Á., Brooks, M. L., Buhrmester, M. D., Vázquez, A., Jetten, J., & Swann, W. B., Jr. (2011). On the nature of identity fusion: Insights into the construct and a new measure. *Journal of Personality and Social Psychology, 100*(5), 918. https://doi.org/10.1037/a0022642.

Gómez, Á., Vázquez, A., López-Rodríguez, L., Talaifar, S., Martínez, M., Buhrmester, M. D., & Swann, W. B., Jr. (2019). Why people abandon groups: Degrading relational vs collective ties uniquely impacts identity fusion and identification. *Journal of Experimental Social Psychology, 85*, 103853.

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin, 135*(4), 555.

Haselmayer, M., Wagner, M., & Meyer, T. M. (2017). Partisan bias in message selection: Media gatekeeping of party press releases. *Political Communication, 34*(3), 367–384.

Hense, R., & Wright, C. (1992). The development of the attitudes toward censorship questionnaire 1. *Journal of Applied Social Psychology, 22*(21), 1666–1675.

Holbert, R. L., Garrett, R. K., & Gleason, L. S. (2010). A new era of minimal effects? A response to Bennett and Iyengar. *Journal of Communication, 60*(1), 15–34.

Iyengar, S., & Hahn, K. S. (2009). Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication, 59*(1), 19–39. https://doi.org/10.1111/j.1460-2466.2008.01402.x.

John, N. A., & Dvir-Gvirsman, S. (2015). "I don't like you any more": Facebook unfriending by Israelis during the Israel–Gaza conflict of 2014. *Journal of Communication, 65*(6), 953–974.

Lawrence, E., Sides, J., & Farrell, H. (2010). Self-segregation or deliberation? Blog readership, participation, and polarization in American politics. *Perspectives on Politics, 8*(1), 141–157.

Leber, c. (2016, January 14). *Gun control can swing the 2016 election.* The New Republic. Retrieved from https://newrepublic.com.

Linder, M. (2016, November 9). Block. Mute. Unfriend. Tensions rise on Facebook after election results. Chicago Tribune. Retrieved from https://www.chicagotribune.com.

Lindner, N. M., & Nosek, B. A. (2009). Alienable speech: Ideological variations in the application of free-speech principles. *Political Psychology, 30*(1), 67–92.

Matias, J. N. (2016a). The civic labor of online moderators. *Internet politics and policy conference*. Oxford: United Kingdom.

Matias, J. N. (2016b). Going dark: Social factors in collective action against platform operators in the Reddit blackout. *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 1138–1151). .

McAdams, D. P. (1995). What do we know when we know a person? *Journal of Personality, 63*(3), 365–396.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27*(1), 415–444.

Motyl, M., Iyer, R., Oishi, S., Trawalter, S., & Nosek, B. A. (2014). How ideological migration geographically segregates groups. *Journal of Experimental Social Psychology, 51*, 1–14.

Muthén, L. K., & Muthén, B. O. (2012). *MPlus: Statistical analysis with latent variables—User's guide*.

Paluck, E. L., Green, S. A., & Green, D. (2018). The contact hypothesis re-evaluated. *Behavioural Public Policy*, 1–30. https://doi.org/10.1017/bpp.2018.25.

Price, V., Nir, L., & Cappella, J. N. (2006). Normative and informational influences in online political discussions. *Communication Theory, 16*(1), 47–74.

Rentfrow, P. J., Gosling, S. D., & Potter, J. (2008). A theory of the emergence, persistence, and expression of geographic variation in psychological characteristics. *Perspectives on Psychological Science, 3*(5), 339–369. https://doi.org/10.1111/j.1745-6924.2008.00084.x.

Riffkin, R. (2015, May 29). Abortion edges up as important voting issue for Americans. Gallup. Retrieved from http://news.gallup.com.

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology, 13*(3), 279–301. https://doi.org/10.1016/0022-1031(77)90049-X.

Schmidt, G. B. (2015). Fifty days an MTurk worker: The social and motivational context for Amazon Mechanical Turk workers. *Industrial and Organizational Psychology, 8*(2), 165–171. https://doi.org/10.1017/iop.2015.20.

Sears, D. O., & Freedman, J. L. (1967). Selective exposure to information: A critical review. *Public Opinion Quarterly, 31*(2), 194–213.

Shaw, A., & Hill, B. M. (2014). Laboratories of oligarchy? How the iron law extends to peer production. *Journal of Communication, 64*(2), 215–238.

Sibona, C. (2014, January). Unfriending on Facebook: Context collapse and unfriending behaviors. *2014 47th Hawaii international conference on system sciences* (pp. 1676–1685). IEEE.

Singh, R., & Ho, S. Y. (2000). Attitudes and attraction: A new test of the attraction, repulsion and similarity-dissimilarity asymmetry hypotheses. *British Journal of Social Psychology, 39*, 197–211.

Singh, R., & Teoh, J. B. P. (1999). Attitudes and attraction: A test of two hypotheses for the similarity-dissimilarity asymmetry. *British Journal of Social Psychology, 38*, 427–443.

Skitka, L. J., & Morgan, G. S. (2014). The social and political implications of moral conviction. In H. Lavine (Vol. Ed.), *Advances in Political Psychology*. *35*. *Advances in Political Psychology* (pp. 95–110).

Skitka, L. J., & Mullen, E. (2002). Understanding judgments of fairness in a real-world political context: A test of the value protection model of justice reasoning. *Personality and Social Psychology Bulletin, 28*(10), 1419–1429.

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology, 88*(6), 895.

Stoycheff, E. (2016). Please participate in part 2: Maximizing response rates in longitudinal MTurk designs. *Methodological Innovations, 9*. https://doi.org/10.1177/2059799116672879 2059799116672879.

Stroud, N. J. (2017). Selective exposure theories. *The Oxford handbook of political communication*.

Suedfeld, P., Steel, G. D., & Schmidt, P. W. (1994). Political ideology and attitudes toward censorship 1. *Journal of Applied Social Psychology, 24*(9), 765–781.

Swann, W. B., Jr., Gómez, A., Seyle, D. C., Morales, J., & Huici, C. (2009). Identity fusion: The interplay of personal and social identities in extreme group behavior. *Journal of Personality and Social Psychology, 96*(5), 995.

Swann, W. B., Jr., Jetten, J., Gómez, Á., Whitehouse, H., & Bastian, B. (2012). When group membership gets personal: A theory of identity fusion. *Psychological Review, 119*(3), 441. https://doi.org/10.1037/a0028589.

Swann, W. B., Jr., Buhrmester, M. D., Gómez, A., Jetten, J., Bastian, B., Vázquez, A., & Finchilescu, G. (2014). What makes a group worth dying for? Identity fusion fosters perception of familial ties, promoting self-sacrifice. *Journal of Personality and Social Psychology, 106*(6), 912.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. *Organizational Identity: A Reader, 56*, 65.

Talaifar, S., & Swann, W. B., Jr. (2019). Deep alignment with country shrinks the moral gap between conservatives and liberals. *Political Psychology, 40*(3), 657–675.

Thomas, E. F., McGarty, C., Reese, G., Berndsen, M., & Bliuc, A. M. (2016). Where there is a (collective) will, there are (effective) ways: Integrating individual-and group-level factors in explaining humanitarian collective action. *Personality and Social Psychology Bulletin, 42*(12), 1678–1692.

TurkPrime (2018, September 18). After the bot scare: Understanding What's been happening with data collection on MTurk and how to stop it [web log post]. Retrieved from https://blog.turkprime.com.

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in Cognitive Sciences, 22*(3), 213–224.

van der Linden, S. (2017). The nature of viral altruism and how to make it stick. *Nature Human Behavior, 1*, 0041. https://doi.org/10.1038/s41562-016-0041.

Van Zomeren, M., Postmes, T., & Spears, R. (2012). On conviction's collective consequences: Integrating moral conviction with the social identity model of collective action. *British Journal of Social Psychology, 51*(1), 52–71.

Wright, S. (2006). Government-run online discussion fora: Moderation, censorship and the shadow of Control1. *The British Journal of Politics and International Relations, 8*(4), 550–568.

Yang, Y. (2019). When power goes wild online: How did a voluntary moderator's abuse of power affect an online community? *Proceedings of the Association for Information Science and Technology, 56*(1), 504–508.

Zaal, M. P., Saab, R., O'Brien, K., Jeffries, C., Barreto, M., & van Laar, C. (2017). You're either with us or against us! Moral conviction determines how the politicized distinguish friend from foe. *Group Processes & Intergroup Relations, 20*(4), 519–539.