Northwestern | INSTITUTE FOR POLICY RESEARCH

IPR Working Paper Series

WP-23-29

The Social Media Discourse of Engaged Partisans Is Toxic Even When Politics Are Irrelevant

Michalis Mamakos
Northwestern University

Eli Finkel
Northwestern University and IPR

Version: July 19, 2023

DRAFT

Please do not quote or distribute without permission.

Abstract

Prevailing theories of partisan incivility on social media suggest that it derives from disagreement about political issues or from status competition between groups. The present study—which analyzes commenting behavior of Reddit users across diverse cultural contexts (subreddits)—tests the alternative hypothesis that such incivility derives in large part from a selection effect: Toxic people are especially likely to opt into discourse in partisan contexts. First, the authors examined commenting behavior across over 9,000 unique cultural contexts (subreddits) and confirmed that discourse is indeed more toxic in partisan (e.g., *r/The_Donald*, *r/LGBTNews*) than in non-partisan contexts (e.g., *r/movies*, *r/programming*). Next, they analyzed hundreds of millions of comments from over 6.3 million users and found robust evidence that: (1) The discourse of people whose behavior is especially toxic in partisan contexts is also especially toxic in non-partisan contexts (i.e., people are not politics-only toxicity specialists); and (2) when considering only non-partisan contexts, the discourse of people who also comment in partisan contexts is more toxic than the discourse of people who do not. These effects were not driven by socialization processes whereby people overgeneralized toxic behavioral norms they had learned in partisan contexts. In contrast to speculation about the need for partisans to engage beyond their echo chambers, toxicity in non-partisan contexts was higher among people who also comment in both left-wing and right-wing contexts (bilaterally engaged users) than among people who also comment in only left-wing or right-wing contexts (unilaterally engaged users). Discussion considers implications for democratic functioning and theories of polarization.

Significance Statement

Political discourse on social media is infamously uncivil. Prevailing explanations argue that such incivility is driven by differences in ideological or social-identity conflict—partisans are uncivil because the political stakes are so high. The present report considers a different (albeit not contradictory) possibility—that online political discourse tends to be uncivil because the people who opt into such discourse are generally uncivil. Indeed, people who opt into political discourse tend to be especially toxic, *even when discussing non-political topics in non-partisan contexts*. Such individuals disproportionately dominate political discourse online, thereby undermining the public sphere as a venue for inclusive debate.

The Social Media Discourse of Engaged Partisans is Toxic Even when Politics are Irrelevant

Partisan hatred is surging, both in the United States (Iyengar et al., 2019; Finkel et al., 2020) and in many other nations (Reiljan, 2020; Wagner, 2021). Such hatred, along with the associated anger, is linked to incivility toward opposing partisans (Webster, Connors, & Sinclair, 2022), especially among those who are deeply engaged in politics (Baldassarri & Gelman, 2008; Rogowski & Sutherland, 2016; Krupnikov & Ryan, 2022). Indeed, animosity toward opposing partisans motivates political engagement on social media (Rathje et al., 2021), where engaged partisans are especially likely to amplify moralized-emotional political content (Brady et al., 2017).

Why are such deeply engaged partisans so uncivil in their political discourse? Two theories prevail. The first focuses on political ideology, suggesting that the politically engaged are especially uncivil because opposing partisans hold attitudes, values, and policy preferences that misalign in important ways with their own (Carmines & Stimson, 1980; Rogowski & Sutherland, 2016; Webster, & Abramowitz, 2017). The second focuses on social identity, suggesting that the politically engaged are especially likely to perceive their group as competing against opposing partisans for resources and status (Green, Palmquist, & Shickler, 2002; Huddy, 2001; Mason, 2018; Van Bavel & Pereira, 2018).

Both of these theories argue that the political context is, for ideological or identarian reasons, a necessary condition for explaining the political incivility of engaged partisans. Partisanship should be irrelevant to incivility in contexts that are irrelevant to politics—contexts in which people gather to discuss, for example, movies, parenting, or programming.

The present report considers a different (albeit not contradictory) possibility, which we call the *troll hypothesis*: that online political discourse tends to be uncivil because the people who opt

into such discourse are generally uncivil. Indeed, people who are more dispositionally disagreeable hold more negative views of opposing partisans (Webster, 2018); those who are more dispositionally aggressive engage in more aggressive political behavior and hold more violent partisan views (Kalmoe & Mason, 2022). Recent articles have demonstrated consistency in hostile behavior in online and offline political discourse (Bor & Petersen, 2022), and that when prompted to comment on posts related to politics, people who have online political activity are more likely to exhibit toxic behavior than people who do not (Kim et al., 2021). However, no research has investigated (1) within-person differences in hostility between partisan and non-partisan contexts or (2) between-person differences in hostility between engaged partisans and the non-engaged in contexts in which politics are irrelevant. Insofar as the incivility of engaged partisans results from broader dispositional tendencies to seek conflict, such individuals are hypothesized to be uncivil in both partisan and non-partisan contexts—and more uncivil than the non-engaged, *even when discussing non-political topics in non-partisan contexts*.

A compelling test of the troll hypothesis requires a study that affords two crucial comparisons. The first compares the behavior of engaged partisans in partisan vs. non-partisan contexts to test whether people are *toxicity specialists* (i.e., only when politics are relevant) vs. *toxicity generalists* (i.e., in both political and non-political contexts). The second compares the behavior of the engaged and the non-engaged in non-partisan contexts to test whether engaged partisans are more toxic than the non-engaged when politics are irrelevant. Ideally, such a study would investigate not one or two of each type of context (partisan and non-partisan), but thousands of them—and those contexts would be highly diverse in terms of their subject matter.

Ideally, the study would also investigate such behavior in an important *public square*—a place where millions or billions of people come to introduce and debate societally important ideas. And it would include both left-wing and right-wing cultural contexts to allow us to explore

whether incivility in non-partisan contexts varies as a function of whether engaged partisans comment on one side vs. both sides of the partisan divide (unilaterally vs. bilaterally engaged partisans). Scholars and social commentators have argued that a major cause of partisan toxicity is the emergence of an "echo chamber" phenomenon in which people encounter people and ideas that come disproportionately from their own side of the divide (e.g., Barberá et al., 2015; Colleoni, Rozza, & Arvidsson, 2014; Sunstein, 2018). However, a major study demonstrated that the political extremity of American partisans actually increased after people were assigned to see social media posts from opposing partisans (Bail et al., 2018). In our study, participants were not randomly assigned to see posts from opposing partisans; rather they had the option of engaging in communities on one side vs. on both sides of the political divide. If the echo chambers hypothesis applies here, then the bilaterals should be less toxic than the unilaterals. In contrast, if the troll hypothesis applies here, then bilaterals should be more toxic, as dispositionally uncivil people are hypothesized to opt into political discourse—to jump into the fray—across the partisan divide.

To meet these criteria, we studied commenting behavior on Reddit from 2011 through 2022. Billions of people around the world use Reddit, which is also the fifth most-visited website in the United States, where it had 2.32 billion visits in March of 2023 alone.[1] Compared to Facebook and Twitter, Reddit is much less dependent on algorithms that determine which information users are exposed to (Waller & Anderson, 2021), which means that behavior on the platform is driven by user decisions to opt in to a given context to make comments rather than being exposed to some contexts rather than others.

We began by considering which cultural contexts (subreddits) are political and which are non-political. Politics can be relevant even in contexts that are not explicitly political, especially

[1] https://www.semrush.com/blog/most-visited-websites/

insofar as groups consisting of politically like-minded people adopt a worldview or style of discourse that leans left or right. Consequently, we employed both a *content* criterion and a *partisan segregation* criterion to establish a given context as *non-partisan*: it must (1) focus on non-political content and (2) be populated about equally by people who tend to lean left vs. right. We operationalized the *partisan segregation* of each subreddit in terms of the extent to which the social networks of contributors to that subreddit overlapped with the contributors in left-wing vs. right-wing political subreddits (Waller & Anderson, 2021). Some highly segregated subreddits are explicitly political (e.g., *r/hillaryclinton*, *r/The_Donald*), whereas others are ostensibly non-political (e.g., *r/librarians*, *r/wrestling*)—but all of them are populated disproportionately with people who generally engage in either left-wing or right-wing social contexts. In this report, we define *engaged partisans* as users with activity in highly segregated subreddits, which may or may not be of explicitly political content.

Results

In our first analysis, we assessed whether users' commenting behavior is indeed more toxic in subreddits that are higher (vs. lower) in partisan segregation, operationalizing *toxicity* using Google's PerspectiveAPI classifier, which assesses the probability that a comment is "rude, disrespectful, or unreasonable and is likely to make someone leave a discussion" (Wulczyn, Thain, & Dixon, 2017). Complementing research demonstrating that social-media discourse is more uncivil in contexts focusing on political than on non-political content (Sun et al., 2021), we tested whether such discourse is more toxic in contexts disproportionately populated by partisans on one side of the political divide (regardless of the contexts' content focus). A random sample of over 260 million comments from 9,364 subreddits (the substantially active of the 10,006 subreddits considered by Waller and Anderson, 2021) revealed a quadratic effect of partisan segregation on toxicity ($\beta = .21$, $p < .0001$). As hypothesized, the association of segregation with toxicity became

increasingly positive at higher levels of segregation. As depicted in Figure 1, partisan segregation and toxicity were largely unrelated in subreddits where segregation is modest, but these two variables were robustly linked in highly segregated subreddits. For example, for subreddits that are at least 2 *SD*s from the neutral point of 0, $r = .25$, $p < .0001$.

Such findings are consistent both with prevailing theories (that partisan incivility on social media results from division across ideology or social identity) and with our troll hypothesis (that people who generally behave toxically are especially likely to opt into partisan contexts). But only the troll hypothesis predicts that engaged partisans are toxicity-generalists whose behavior is uncivil even in contexts that are non-partisan and non-political. As a first test of this idea, we classified as non-partisan those subreddits with partisan segregation scores within 0.25 *SD*s from the neutral point of 0 ($N_{NonpartisanSubreddits} = 2{,}084$), and as partisan those subreddits with partisan segregation scores at least 2 *SD*s away from that neutral point ($N_{PartisanSubreddits} = 467$).[2] We analyzed toxicity for users who made at least five comments both in partisan and in non-partisan subreddits within a year of their registration on Reddit ($N_{Engaged} = 1{,}045{,}631$), excluding comments in non-partisan subreddits that were classified as political comments (based on the dictionary of Simchon, Brady, and Van Bavel, 2022). In support of the troll hypothesis, Figure 2 reveals that the toxicity these users exhibited in partisan subreddits was highly correlated with their toxicity in non-partisan subreddits ($r = .47$). An auxiliary analysis studying only those users who commented at least 20 times each in partisan and non-partisan subreddits (i.e., those users for whom we have an especially reliable measure of toxicity) suggests that the actual correlation may be even higher ($r = .60$). In short, people are toxic in partisan contexts in large part because they are toxic in general.

[2] To meet our inclusion criteria for establishing a context as non-partisan, we excluded 16 non-partisan subreddits that were explicitly political (0.76% of the non-partisan subreddits). For the partisan subreddits, 105 (22.48%) were explicitly political; later, we report results separately for partisan subreddits of political vs. non-political content.

As a second test, we focused exclusively on non-partisan contexts, comparing the commenting behavior of these engaged partisans with that of the non-engaged—users who made at least five comments in non-partisan subreddits but none in partisan subreddits ($N_{NonEngaged}$ = 5,255,708). For this comparison, we divided the engaged users into two subgroups: (1) the *unilaterally engaged*, who commented in only left-wing or only right-wing partisan subreddits ($N_{Unilaterals}$ = 681,311; 57% were left-wing only); and (2) the *bilaterally engaged*, who commented in both left-wing and right-wing subreddits ($N_{Bilaterals}$ = 364,320).

Figure 3 depicts the toxicity of these three groups in non-partisan subreddits. Relative to the commenting behavior of the non-engaged (Figure 3a, green violin plot on the left), the commenting behavior of the unilaterally engaged (Figure 3a, orange violin plot in the middle) was substantially more toxic (d = 0.26). Robustness checks revealed that this effect also emerged for auxiliary measures of incivility (Figure S1 in *SI Appendix*): Relative to the non-engaged, the unilaterally engaged expressed greater moral outrage (d = 0.21) and were less polite (d = –0.16) and less prosocial (d = –0.17). They were also more profane (d = 0.08) and more angry (d = 0.09), although those effects are small. In short, when discussing non-political topics in non-partisan subreddits, the commenting behavior of unilaterally engaged partisans is more uncivil than that of the non-engaged.

What about the bilaterally engaged? Here we consider competing hypotheses. Insofar as toxicity is caused in part by echo-chamber dynamics that prevent social media users from engaging with opposing partisans, the unilaterally engaged might be more toxic than the bilaterally engaged (*the echo chambers hypothesis*). Alternatively, insofar as people who are generally inclined to engage in toxic discourse seek out highly partisan contexts across the political spectrum, the bilaterally engaged might be even more toxic than the unilaterally engaged (*the bilateral troll hypothesis*).

The results presented in Figure 3a disconfirm the echo chambers hypothesis and support the bilateral troll hypothesis. Bilaterally engaged partisans (Figure 3a, purple violin plot on the right) were more toxic than the unilaterally engaged (d = 0.28) and far more toxic than the non-engaged (d = 0.54). Robustness checks revealed that this tendency for bilaterally engaged partisans to be more toxic than the non-engaged also emerged for the auxiliary measures of incivility (Figure S1 in *SI Appendix*): Relative to the non-engaged, the bilaterally engaged expressed greater moral outrage (d = 0.36), were less polite (d = –0.29), and were less prosocial (d = –0.31). They were also more profane (d = 0.20) and more angry (d = 0.15).

The results in Figure 3a, which emerge across all cohorts of Reddit registrants (see Figure 3b), provide support for the troll and bilateral troll hypotheses: that engaged partisans (especially the bilaterally engaged) are more uncivil than the non-engaged, even when politics are irrelevant. We subjected these findings to five robustness checks. First, perhaps the results are not about incivility in particular, but about *negativity in general*, including the "internalizing" tendencies of anxiety and sadness (Ekman, 1993; Lerner & Keltner, 2000; Smith & Kirby, 2004). However, we find that the levels of anxiety and sadness expressed in the comments were nearly identical across the non-engaged, the unilaterally engaged, and the bilaterally engaged (all ds < 0.04).

Second, perhaps the toxic behavior of engaged partisans in non-partisan subreddits results not from a dispositional tendency toward incivility but rather from a socialization process in which engagement in partisan subreddits teaches them uncivil norms, which they then overgeneralize to non-partisan subreddits (*the socialization hypothesis*). To explore this possibility, we conducted a longitudinal analysis of the users who exhibited partisan engagement. We modeled the toxicity of the comments these users made in non-partisan subreddits as a function of the partisan activity those users had by the time of posting. A fixed-effects (within) estimator revealed that partisan activity effectively explains 0% of the variance (R^2 < .001) of toxicity in non-partisan subreddits.

Third, perhaps the results in Figure 3 are driven only by users whose engagement in highly segregated subreddits is limited to subreddits of *political content* (e.g., *r/hillaryclinton*, *r/The_Donald*)—or, alternatively, to subreddits that are ostensibly non-political (e.g., *r/librarians*, *r/wrestling*). To consider this possibility, we split the engaged into two groups: those with vs. without any comments in partisan subreddits of political content. As illustrated in Figure 4, the tendency of the unilaterally engaged and (especially) the bilaterally engaged to be more toxic in non-partisan subreddits emerged regardless of whether partisans also engage in partisan subreddits that were explicitly political or ostensibly non-political, but the effects were especially strong for partisans who also engaged in partisan subreddits that were explicitly political. The effect sizes for explicitly political vs. ostensibly non-political subreddits were $d = 0.43$ vs. $d = 0.20$ for the unilaterals and $d = 0.62$ vs. $d = 0.36$ for the bilaterals.

Fourth, we examined whether the observed differences in toxicity are moderated by the *political lean* of the engaged. The users whose partisan engagement was only with left-wing subreddits were virtually exactly as toxic as those whose partisan engagement was only with right-wing subreddits ($d = 0.27$ and $d = 0.25$, respectively, compared to the non-engaged). The bilaterally engaged who commented predominantly in left-wing subreddits (47% of the bilaterally engaged) also exhibit virtually the same level of toxicity as their right-wing counterparts ($d = 0.56$ and $d = 0.53$, respectively, compared to the non-engaged).

And fifth, perhaps the Figure 3 results were driven by behavior in a small number of *outlier subreddits*, albeit perhaps highly populated ones. To explore this possibility, we considered the 1,221 non-partisan subreddits in which at least 1,000 comments were posted by each of the three groups (non-partisans, unilaterals, and bilaterals). The unilaterally engaged were more toxic than the non-engaged in 97% of those subreddits, and the bilaterally engaged are more toxic in 99% of them. We created a subreddit-specific toxicity ratio of the comments made by engaged partisans to

the comments made by the non-engaged. Figure S4 in *SI Appendix* presents a histogram of the results for unilaterals and bilaterals, demonstrating that the average subreddit exhibits a 13.2% toxicity increase for unilaterally engaged partisans relative to the non-engaged (95% confidence interval: 12.7%-13.8%) and a 25.6% toxicity increase for bilaterally engaged partisans (95% confidence interval: 24.8%-26.5%).

Discussion

Taken together, the results provide strong and consistent support for the troll hypothesis: (1) people who are especially toxic in partisan contexts are also especially toxic in non-partisan contexts (Figure 2), and (2) engaged partisans (especially the bilaterally engaged) are more toxic than the non-engaged when discussing non-political content in non-partisan contexts (Figures 3 and 4). Such effects are specific to uncivil behaviors (rather than to negativity in general) and do not result from some sort of socialization process in partisan subreddits. They emerge regardless of political lean, and they apply to users whose partisan comments take place in contexts that are explicitly political or ostensibly non-political—although they are especially strong for the users with activity in explicitly political contexts. The effects, which emerge in virtually all non-partisan subreddits, help to explain why political contexts tend to be more toxic than non-political contexts (Figure 1).

Future research will be required to test how strongly these results generalize beyond Reddit. That said, a strength of the present study is that it investigates hundreds of millions of unique behaviors from millions of people across thousands of cultural contexts (subreddits). As such, the results are not subject to the typical concerns about a limited range of cultures or topics of discourse. In addition, social media environments (e.g., Twitter, Facebook, Reddit) have become a core nexus for political discourse, increasingly functioning as democracy's public square (Van Bavel et al., 2021). Reddit is a major context where political ideas get introduced and debated—

where people of diverse backgrounds and ideologies discuss and argue about which ideas and policies are best (Hofmann, Schütze, & Pierrehumbert, 2022).

The present findings have important implications for theories of political polarization. They suggest that discourse in partisan contexts is uncivil in large part because the people who opt into it are uncivil. This incivility distorts the public square. People's reluctance to contribute to political discourse—to contribute their views to the marketplace of ideas—is driven less by substantive disagreement than by the tenor of the discourse; they opt out when discourse gets heated (Connors & Howell, 2022; Sydnor, 2019). It is no wonder that people who are lower in trait hostility tend to opt out of online political discourse (Bor & Petersen, 2022). The overrepresentation of dispositionally uncivil people in our political discourse is especially troubling because it promotes combative partisanship at the expense of deliberation (Gervais, 2019) and leads observers (those who also participate and those who do not) to conclude that the state of our politics is far more toxic than it really is (Brady et al., 2023).

There is little reason to believe that dispositionally uncivil people have better political ideas than those who are more dispositionally civil, and there is good reason to believe that the uncivil are less prone to compromise, to seek win-win solutions, or to assume that their interlocutors are people of goodwill (Krupnikov & Ryan, 2022). Consequently, the disproportionate representation of uncivil people in partisan contexts may be a significant contributor to the democratic backsliding afflicting the United States and many other nations in recent years (Levitsky & Ziblatt, 2019). Theories of polarization must engage seriously with the fact that society has built a new megaphone that amplifies the voices of people whose discourse tendencies are disproportionally characterized by toxicity, moral outrage, profanity, anger, impoliteness, and low prosociality.

Past research has demonstrated that passive exposure to social media posts from opposing partisans can exacerbate polarization (Bail et al., 2018), but the present study is the first to test

whether people who opt into partisan discourse on one vs. both sides of the political divide tend to be especially toxic. Reddit offers to its users the opportunity to join multiple communities across the political spectrum, and it gives space for constructive conversations on controversial topics. Nevertheless, our results suggest that this opportunity is exploited by people with especially uncivil tendencies. These findings contribute to an emerging sense of skepticism about whether breaking down echo chambers will reduce polarization or toxicity—at least in a straightforward way.

Democracy requires conflict. People with differing ideological and policy preferences must compete in the marketplace of political ideas, seeking to persuade others that their own ideas are best. The present research suggests, however, that the voices that are most amplified on social media are dispositionally toxic, an arrangement that seems unlikely to cultivate the sort of constructive discussion and debate that democracies require. Consequently, an urgent priority for societies riven by polarization and democratic backsliding is to develop a means of making the public square a congenial environment not only for the dispositionally uncivil, but also for people who would be willing to enter the debate if only the tenor of the discourse were less toxic.

Materials and Methods

We used the Pushshift Reddit dataset (Baumgartner et al., 2020), which includes information about the comments made on Reddit: the author, the posting date, the subreddit, the content, and the unique identifier of a comment. We excluded comments made by users whose username includes the word "bot" and by moderators. PerspectiveAPI has by default a quota limit of 1 query per second. To analyze millions of comments, we made a request for a limit of 1,000 queries per second. This request was approved for a prespecified, limited period.

Our measure of partisan segregation of the subreddits was the absolute value of partisanship derived for 10,006 subreddits by Waller and Anderson (2021), who examined all comments on Reddit from 2005 to 2018 to derive a network-based characterization of subreddit partisanship, independent of the content of these comments. This measure of partisanship was *z*-scored, so that

the neutral point of 0 corresponded to the average partisanship across the subreddits, and the score of each subreddit was in standard deviation (*SD*) units. The more negative the partisanship of a subreddit, the more left-wing the subreddit, and equivalently for positive-valued (right-wing) subreddits. We categorized subreddits as of either political or non-political content based on the hierarchical clustering for content-based categorization performed in a separate analysis by Waller and Anderson (2021).

In our subreddit-level analysis of the relation between toxicity and partisan segregation (results in Figure 1), we considered the 9,364 (of the 10,006) subreddits in which at least 10,000 comments were posted from 2011 to 2022 (inclusive). Our available computing resources allowed us to randomly sample for these subreddits a total of 260,425,138 comments (M = 27,811, SD = 5,725) from that period. We characterized the toxicity of a subreddit by averaging the toxicity of the comments posted in it. Six subreddits with outlier values were excluded from Figure 1 in the main text to enhance graphical clarity, but no outliers were excluded from the quadratic regression itself.

In our user-level analyses, we considered users who registered on Reddit in the period between 2011 and 2021 (inclusive), and we examined their commenting behavior within a year from their registration (e.g., the commenting behavior of a user who registered on 31 December 2021 would be included through 31 December 2022). The number of users for each cohort is presented in Figure S1 (*SI Appendix*). We discarded cohorts before 2011 because they lacked enough users (fewer than 100,000 users from 2005 up to 2010, combined) who satisfied our inclusion criterion of having at least five comments in non-partisan subreddits. In addition, to address the possibility that some political comments might make their way into subreddits that are both non-political in content and non-partisan in segregation (within 0.25 *SD*s of 0), we discarded all comments in non-partisan subreddits that include words classified as *issue-based political* by the dictionary-based approach of Simchon, Brady, and Van Bavel (2022). For instance, the words "political", "bipartisan", "democrat", "republican", and "amendment" are some of the words included in this publicly available dictionary.

Due to limitations of our computing resources, for users with more than 120 comments in non-partisan subreddits (8% of the users), we randomly sampled 120 of their comments. Similarly, for the users with partisan engagement who have more than 120 comments in partisan subreddits (9% of engaged), we randomly sampled 120 of their comments in these contexts. The toxicity of a user was derived by averaging the toxicity of the user's comments. Of the 1,045,631 engaged

users, 310,830 (30%) made at least 20 comments in both partisan and non-partisan subreddits (these users are included in the reported auxiliary analysis about the within-subject correlation of the engaged).

In the model developed for the second robustness check (testing the socialization hypothesis), we included two binary predictors, whether the user already had (a) unilateral and (b) bilateral partisan engagement, and two continuous predictors, the number of comments in (c) left-wing and in (d) right-wing partisan subreddits the user had made by the time of comment-posting in a non-partisan subreddit. Because the number of comments in partisan subreddits can vary greatly across the engaged ($M = 54$, $SD = 204$), but the toxicity of these users in non-partisan subreddits is much more concentrated ($M = 0.15$, $SD = 0.08$), it was plausible that our model's dependent variable might have a sub-linear dependence on the continuous predictors. To account for this possibility, we also included in our model lower order terms of each of the two continuous predictors, ranging from their second (square) to their tenth root, building a model with 22 predictors in total (2 binary and 20 continuous). In the fifth robustness check, the 95% confidence intervals about the subreddit toxicity increase for the users with partisan engagement are bootstrapped (10,000 repetitions).

In addition to the toxicity of PerspectiveAPI, we also assessed several additional measures that are also arguably proxies for incivility. The *moral outrage* of the comments is assessed with the classifier of Brady et al. (2021). This classifier assesses the probability that a comment expresses feelings in response to a violation of moral norms, and where these feelings are comprised of emotions such as anger, disgust, and contempt. For *profanity*, *anger*, *politeness*, *prosociality*, *anxiety*, and *sadness*, we employed the dictionary-based approach of the Linguistic Inquiry and Word Count (LIWC; Boyd et al., 2022). Because this approach can be executed only in a centralized fashion, which makes difficult the assessment for comments whose number is in the hundreds of millions, we developed our own dictionary-based method by reverse-engineering LIWC. We purchased a LIWC license and analyzed over 760,000 unique words with that official software. The results in Table S2 (*SI Appendix*) demonstrate that our dictionary method provides a very close approximation of LIWC. We evaluated the comments of the users with our dictionaries, and then characterized the users based on the averages over their comments.

References

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Lee, J, Mann, M., Merhout 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.

Baldassarri, D., & Gelman, A. (2008). Partisans without constraint: Political polarization and trends in American public opinion. *American Journal of Sociology*, *114*(2), 408-446.

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.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020, May). The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 14, pp. 830-839).

Bor, A., & Petersen, M. B. (2022). The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American Political Science Review*, *116*(1), 1-18.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, 1-47.

Brady, W. J., McLoughlin, K. L., Torres, M. P., Luo, K. F., Gendron, M., & Crockett, M. J. (2023). Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility. *Nature Human Behaviour*, 1-11.

Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*(33), eabe5641.

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.

Carmines, E. G., & Stimson, J. A. (1980). The two faces of issue voting. *American Political Science Review*, *74*(1), 78-91.

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication*, *64*(2), 317-332.

Connors, E. C., & Howell, C. (2022). *"You need to calm down": How tone shapes political discussion.* Unpublished manuscript, University of Southern California.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48*(4), 384–392.

Finkel, E. J., Bail, C. A., Cikara, M., Ditto, P. H., Iyengar, S., Klar, S., ... & Druckman, J. N. (2020). Political sectarianism in America. *Science*, *370*(6516), 533-536.

Gervais, B. T. (2019). Rousing the partisan combatant: Elite incivility, anger, and antideliberative attitudes. *Political Psychology*, *40*(3), 637-655.

Green, D. P., Palmquist, B., & Schickler, E. (2004). *Partisan hearts and minds: Political parties and the social identities of voters*. New Haven, CT: Yale University Press.

Hofmann, V., Schütze, H., & Pierrehumbert, J. B. (2022). The Reddit Politosphere: A Large-Scale Text and Network Resource of Online Political Discourse. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 16, pp. 1259-1267).

Huddy, L. (2001). From social to political identity: A critical examination of social identity theory. *Political Psychology*, *22*(1), 127-156.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, *22*, 129-146.

Kalmoe, N. P., & Mason, L. (2022). *Radical American partisanship: Mapping violent hostility, its causes, and the consequences for democracy*. Chicago, IL: University of Chicago Press.

Kim, J. W., Guess, A., Nyhan, B., & Reifler, J. (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication*, *71*(6), 922-946.

Krupnikov, Y., & Ryan, J. B. (2022). *The other divide: Polarization and disengagement in American politics*. New York: Cambridge University Press.

Lerner, J. S., & Keltner, D. (2000). Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition & emotion*, *14*(4), 473-493.

Levitsky, S., & Ziblatt, D. (2018). *How democracies die*. New York: Crown.

Mason, L. (2018). *Uncivil agreement: How politics became our identity*. Chicago, IL: University of Chicago Press.

Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26).

Reiljan, A. (2020). 'Fear and loathing across party lines'(also) in Europe: Affective polarisation in European party systems. *European Journal of Political Research*, *59*(2), 376-396.

Rogowski, J. C., & Sutherland, J. L. (2016). How ideology fuels affective polarization. *Political Behavior*, *38*, 485-508.

Simchon, A., Brady, W. J., & Van Bavel, J. J. (2022). Troll and divide: The language of online polarization. *PNAS nexus*, *1*(1), pgac019.

Smith C. A., & Kirby, L. D. (2004). Appraisal as a Pervasive Determinant of Anger. *Emotion, 4*(2), 133–138.

Sydnor, E. (2019). *Disrespectful democracy: The psychology of political incivility*. Columbia University Press.

Sun, Q., Wojcieszak, M., & Davidson, S. (2021). Over-time trends in incivility on social media: evidence from political, non-political, and mixed sub-reddits over eleven years. *Frontiers in Political Science*, *3*, 741605.

Sunstein, C. R. (2018). *#Republic: Divided democracy in the age of social media*. Princeton University Press.

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 Bavel, J. J., Rathje, S., Harris, E., Robertson, C., & Sternisko, A. (2021). How social media shapes polarization. *Trends in Cognitive Sciences*, *25*(11), 913-916.

Wagner, M. (2021). Affective polarization in multiparty systems. *Electoral Studies*, *69*, 102199.

Waller, I., & Anderson, A. (2021). Quantifying social organization and political polarization in online platforms. *Nature*, *600*(7888), 264-268.

Webster, S. W. (2018). It's personal: The Big Five personality traits and negative partisan affect in polarized US politics. *American Behavioral Scientist*, *62*(1), 127-145.

Webster, S. W., & Abramowitz, A. I. (2017). The ideological foundations of affective polarization in the US electorate. *American Politics Research*, *45*(4), 621-647.

Webster, S. W., Connors, E. C., & Sinclair, B. (2022). The social consequences of political anger. *The Journal of Politics*, *84*(3), 1292-1305.

Wulczyn, E, Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 1391–1399).

Figure 1. The toxicity and partisan segregation of 9,364 subreddits. The color of a dot (blue or red) indicates the partisan lean (left-wing or right-wing) of that subreddit.

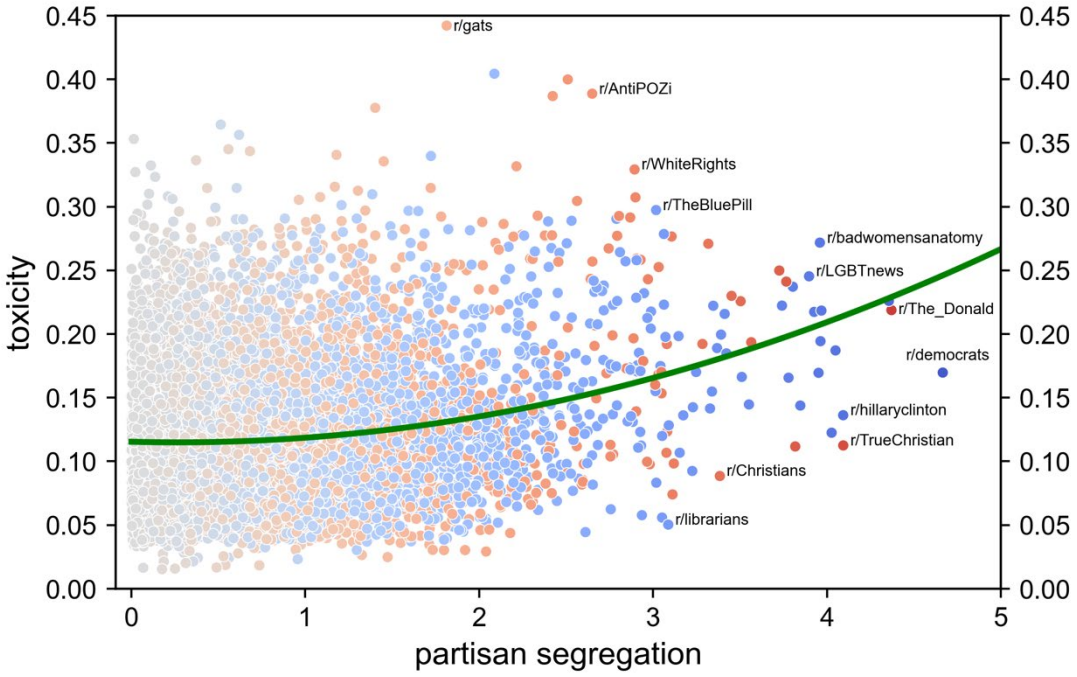


Figure 2. Within-subject correlation of the toxicity of the users with partisan engagement across partisan and non-partisan subreddits. This random sample of 50,000 engaged users exhibited the same correlation as the full sample of the 1,045,631 engaged users (r = .47).

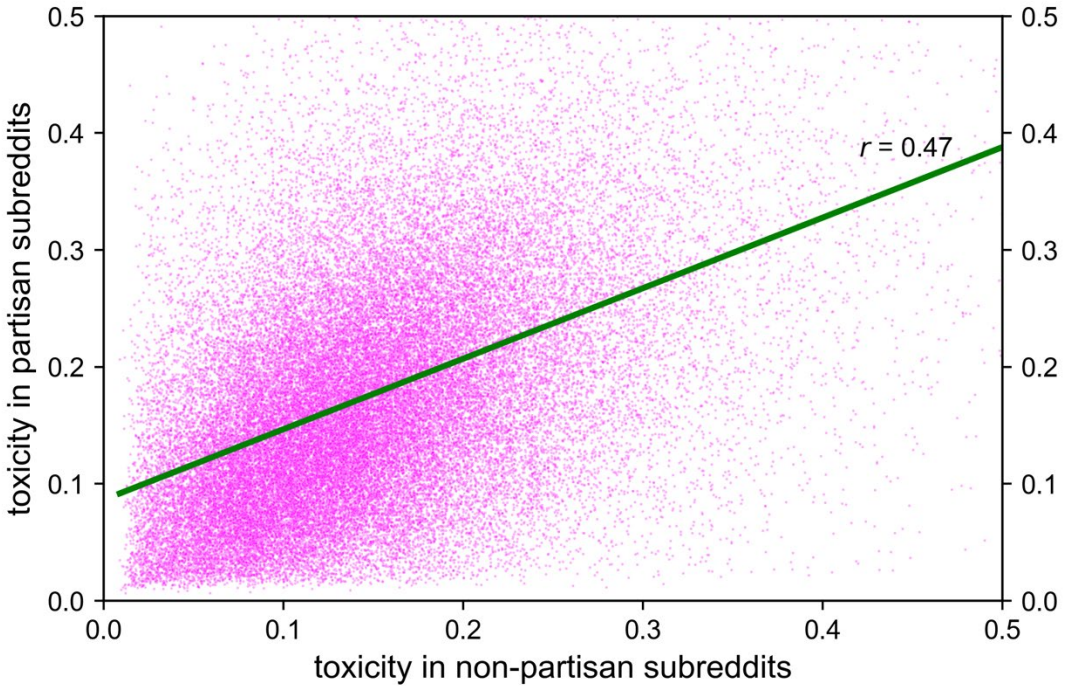
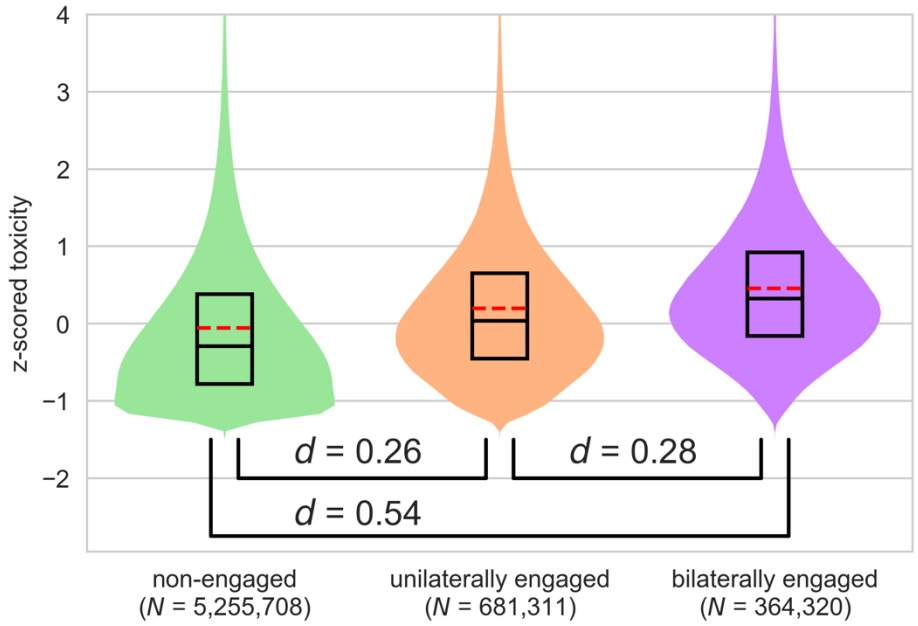
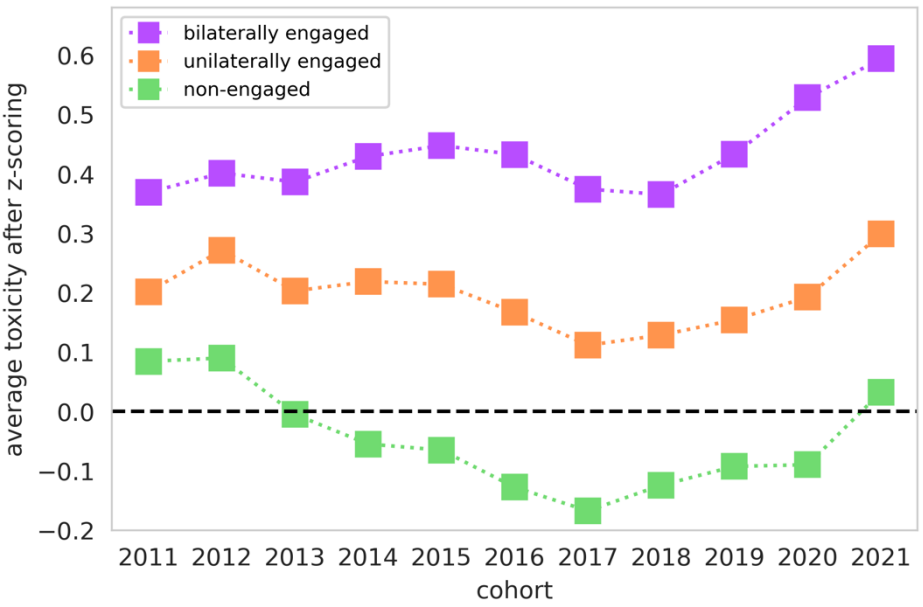


Figure 3. Comparison of the toxicity of the non-engaged, of the unilaterally engaged, and of the bilaterally engaged in non-partisan subreddits.

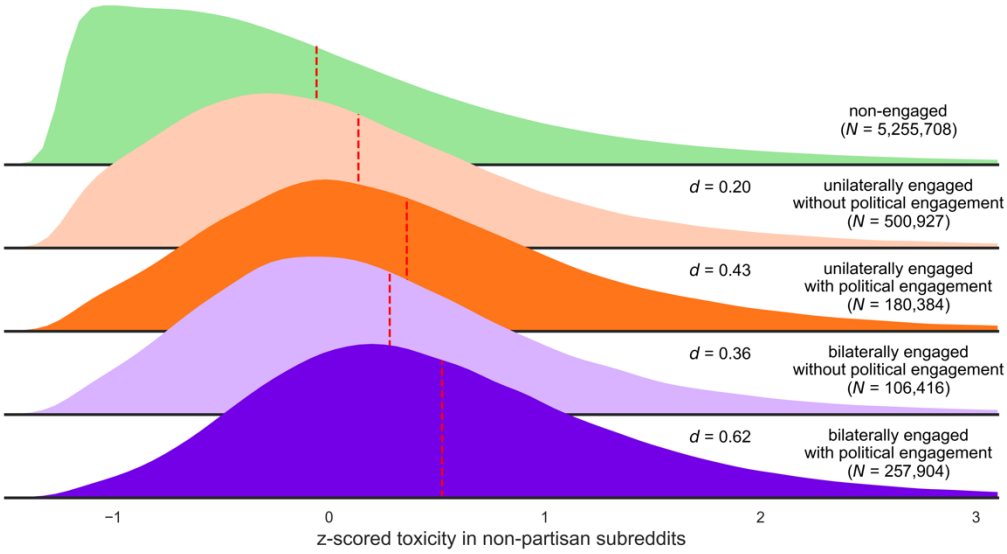


(a) Violin and box plots. The dashed red lines indicate the means.



(b) A cohort corresponds to the year of registration on Reddit.

Figure 4. Density plots about the toxicity of five groups of users in non-partisan subreddits. The dashed red lines indicate the means. Cohen's *d*s are in comparison to the non-engaged.



Data Availability Statement

We commit to making publicly available the data and the code for the analysis of the results by uploading them on GitHub, if this paper is accepted for publication.

The Social Media Discourse of Engaged Partisans is Toxic Even when Politics are Irrelevant

Supplementary Information

Number of users
In Table S1 we present for each cohort the number of users in our sample. The results in Figures 2, 3, and 4 of the main text were derived by analyzing the commenting behavior these users exhibited within a year from their registration on Reddit.

Table S1. The number of users per cohort.

cohort	non-engaged	unilaterally engaged	bilaterally engaged
2011	149,559	16,949	7,915
2012	268,120	34,324	13,336
2013	275,149	30,391	12,338
2014	324,920	34,055	14,047
2015	367,536	35,819	21,392
2016	461,451	49,190	38,329
2017	353,968	48,259	38,212
2018	413,110	76,287	47,244
2019	665,608	107,643	55,329
2020	1,033,400	134,786	66,723
2021	942,887	113,608	49,455
Total N = 6,301,339	5,255,708	681,311	364,320

Auxiliary measures of incivility
The results in Table S2 demonstrate that our dictionary method is a very close approximation of LIWC.

Table S2. The correlation between LIWC and our dictionaries evaluated on a random sample of 20,000 Reddit comments.

profanity	.99
anger	.94
politeness	.95
prosociality	.97
anxiety	.99
sadness	.95

In Figure S1 we present a graphical representation of the results reported in the main text about the auxiliary measures of incivility. In Figure S2, we present the correlations among the different behavioral measures based on the commenting behavior of the users in non-partisan subreddits. Our incivility measure was highly correlated with moral outrage and profanity (r = ~|.5|), moderately correlated with politeness and prosociality (r = ~|.2|), and slightly correlated with anger (r = ~|.1|). As anticipated, it was largely uncorrelated with anxiety and sadness (r = ~|.0|).

Figure S1. Comparison of different measures of incivility of the non-engaged, of the unilaterally engaged, and of the bilaterally engaged in non-partisan subreddits. Positive values of Cohen's *d* correspond to greater values for the users with partisan engagement.

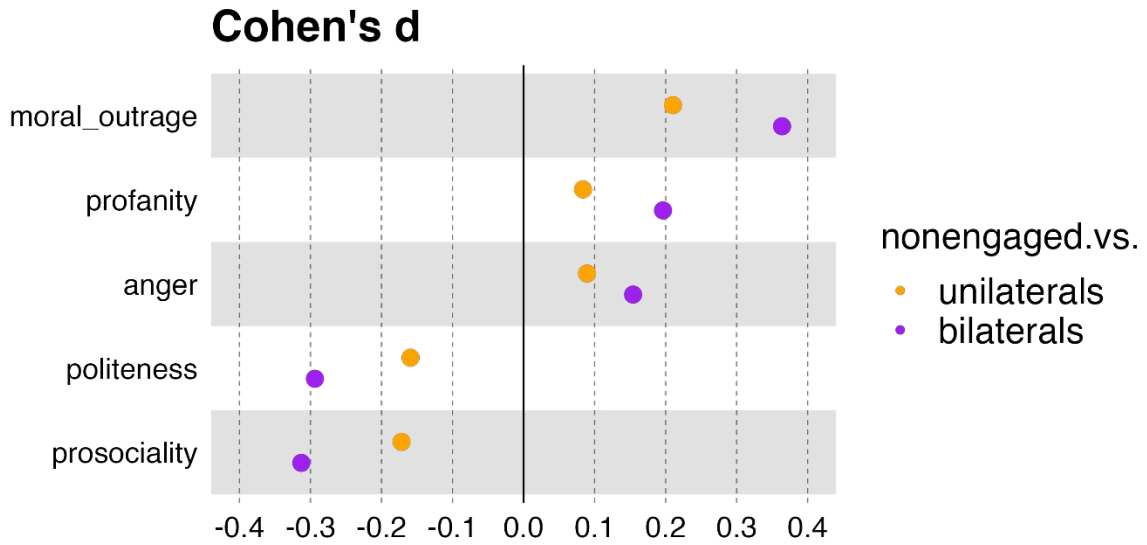
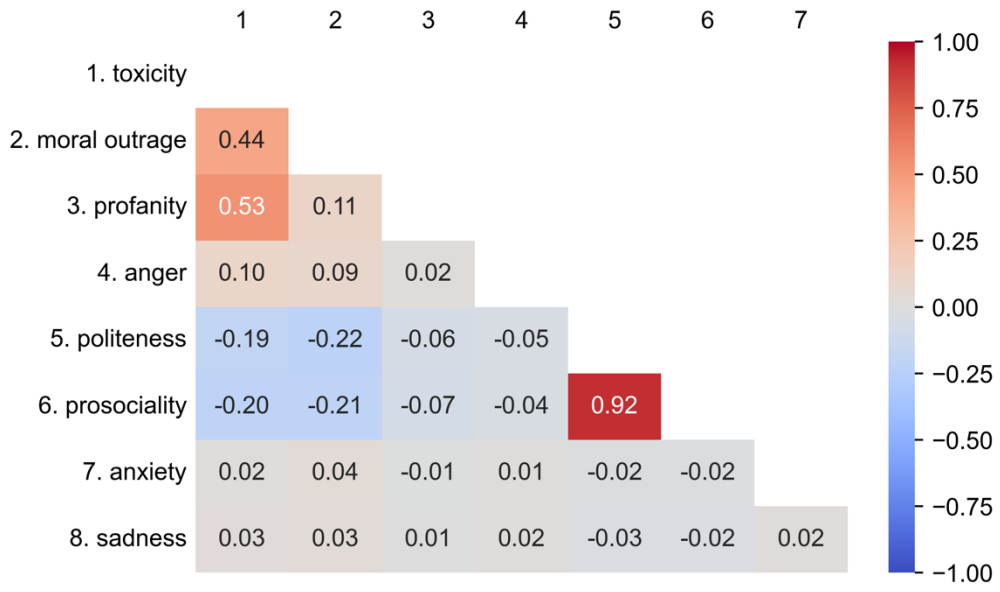


Figure S2. The correlation of behavioral measures as assessed for users based on their comments in non-partisan subreddits.



Number of comments

The analysis in the second robustness check (testing the socialization hypothesis) involved the number of comments that engaged partisans made in partisan subreddits. Here, we first report summary statistics for the number of comments that each group made in non-partisan (Table S3) and, for the engaged, in partisan subreddits (Table S4).

In Table S3 we observe that in non-partisan subreddits, the bilaterally engaged users made on average more comments than the unilaterally engaged users (M = 130 vs. 74), who in turn made more comments than the non-engaged (M = 34). Therefore, it is possible that the heightened toxicity (average toxicity over comments) of the engaged is explained by their high levels of activity in non-partisan subreddits, rather than by the fact that they have opted into partisan contexts. The results in Table S5 reject this possibility. In Column A, the toxicity in non-partisan subreddits was regressed on two binary predictors, one each for unilateral and bilateral partisan engagement. When the number of comments in non-partisan subreddits was also included in the predictors (Column B), the estimates for the coefficients of the two binary predictors changed only marginally, from 0.0226 to 0.0224 for the unilaterals and from 0.0456 to 0.0450 for the bilaterals. A similar observation was made when the number of comments in partisan subreddits was also included in the regression (Column C), as expected based on the results of the second robustness check (which involved only the engaged). These results imply that the increased toxicity the engaged exhibited in non-partisan subreddits is not explained by their number of comments in partisan or non-partisan subreddits. Also, the results in Column A imply that the average toxicity for the non-engaged was 0.121, for the unilaterally engaged 0.143, and for the bilaterally engaged 0.166.

Table S3. Comments in non-partisan subreddits.

	Mean	SD	Median
non-engaged	34	162	12
unilaterally engaged	74	210	26
bilaterally engaged	130	384	53

Table S4. Comments in partisan subreddits.

	Mean	SD	Median
unilaterally engaged	**42**	**159**	**13**
bilaterally engaged	**77**	**268**	**20**

Table S5. Regression results for toxicity in non-partisan subreddits. Controlling for the number of comments does not change the correlation between toxicity and partisan engagement.

	(A)	(B)	(C)
intercept	**0.1211**	**0.1209**	**0.1209**
unilaterally engaged	**0.0226**	**0.0224**	**0.0221**
bilaterally engaged	**0.0456**	**0.0450**	**0.0445**
number of comments in non-partisan subreddits	**-**	**0**	**0**
number of comments in partisan subreddits	**-**	**-**	**0**

Next, we examined whether having just one comment in partisan subreddits was sufficient to predict toxicity in non-partisan subreddits. The results in Figure S3 reveal that this is the case: In predicting behavior in non-partisan subreddits, users with exactly one comment in partisan subreddits were more toxic than those with zero comments in partisan subreddits (*d* = 0.23).

Figure S3. Comparison of toxicity in non-partisan subreddits. The users with exactly one comment are not part of any of our other analyses.

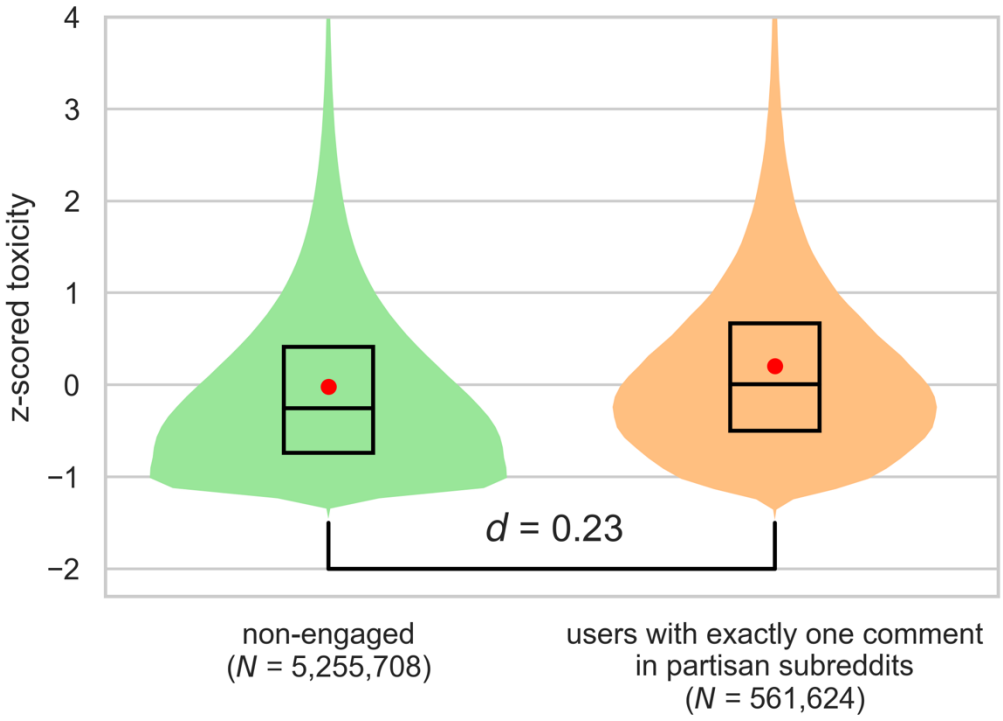
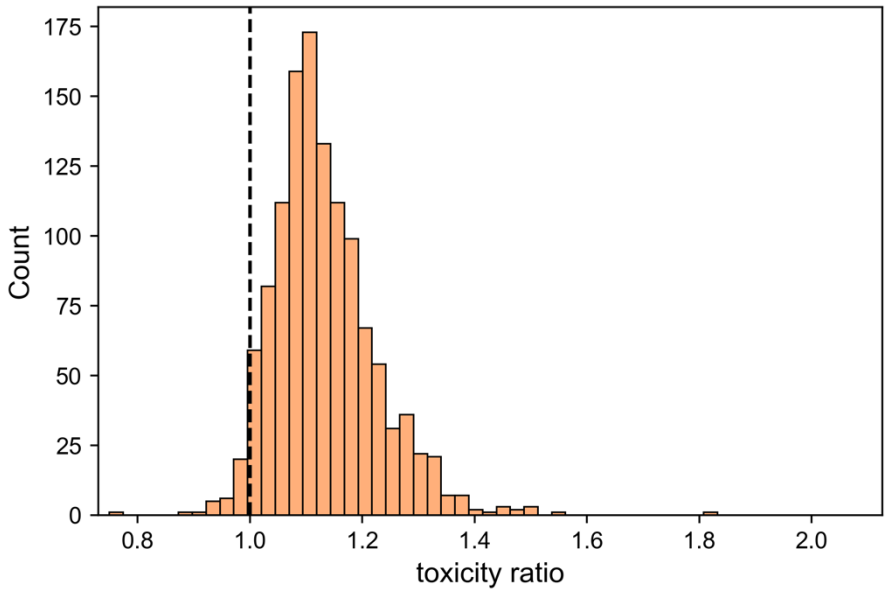
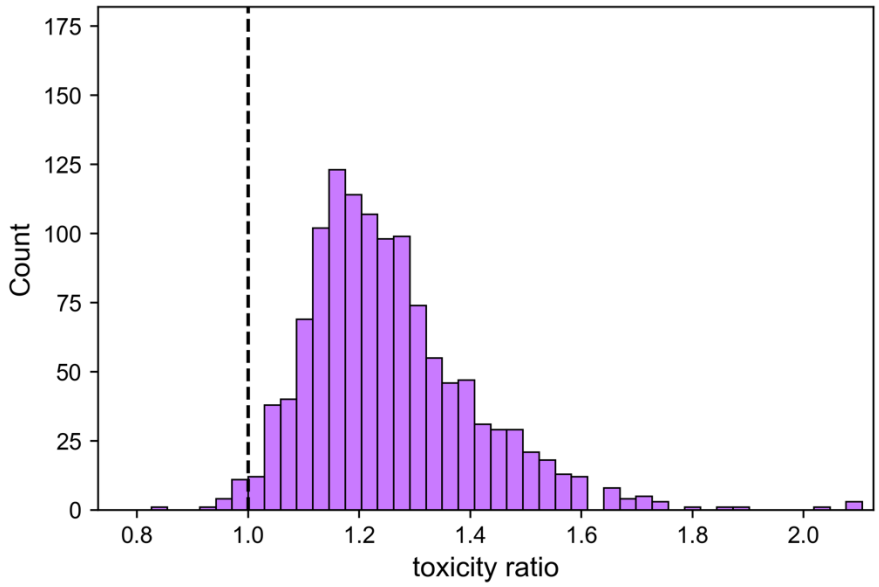


Figure corresponding to the fifth robustness check in the main text
Figure S4. Frequency of non-partisan subreddits based on the toxicity-ratio of the comments made by engaged partisans to the comments made by the non-engaged.



(a) unilaterally engaged vs. non-engaged



(b) bilaterally engaged vs. non-engaged