The Distorting Prism of Social Media
How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity[†]

Forthcoming, *Journal of Communication*

Jin Woo Kim
Annenberg School for Communication
University of Pennsylvania

Andrew Guess
Department of Politics
Princeton University

Brendan Nyhan
Department of Government
Dartmouth College

Jason Reifler
Department of Politics
University of Exeter

Abstract

Though prior studies have analyzed the textual characteristics of online comments about politics, less is known about how selection into commenting behavior and exposure to other people's comments changes the tone and content of political discourse. This article makes three contributions. First, we show that frequent commenters on Facebook are more likely to be interested in politics, to have more polarized opinions, and to use toxic language in comments in an elicitation task. Second, we find that people who comment on articles in the real world use more toxic language on average than the public as a whole; levels of toxicity in comments scraped from media outlet Facebook pages greatly exceed what is observed in comments we elicit on the same articles from a nationally representative sample. Finally, we demonstrate experimentally that exposure to toxic language in comments increases the toxicity of subsequent comments.

Keywords— Social media, Online comments, Toxicity, Self-selection, Polarization

[†]We thank Dartmouth College and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 682758) for funding support. We are grateful to Sam Luks and Marissa Shih at YouGov for survey assistance. We also thank Rachel Bernhard, George Berry, Nathan Kalmoe, Devra Moehler, and workshop and seminar participants at Princeton University, the University of Zurich, and the Research on Online Political Hostility (ROPH'20) conference at Aarhus University for helpful feedback. Guess and Nyhan have received funding from Facebook for other projects, but Facebook played no role in this research. All conclusions and any errors are our own.

The Distorting Prism of Social Media

How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity

Although Americans report overwhelming agreement that the public is sharply divided along partisan lines (Pew Research Center 2019), the best available evidence suggests that partisans exaggerate the extremism of the other side (Ahler 2014; Levendusky & Malhotra 2015; Mason 2018). One important source of these perceptual errors may be what we see online and on social media (Settle 2018; Suhay et al. 2018; Yang et al. 2016). Just as the most committed ideologues are more likely to run for office (Hall 2019), people who discuss politics online are likely to have stronger partisan attachments than the average citizen.

In particular, the skew in who participates in online discussions about politics may contribute to incivility in online comments (e.g., Coe et al. 2014; Muddiman & Stroud 2017). These more extreme comments may in turn inflame polarization (e.g., Suhay et al. 2018) and serve as unrepresentative exemplars that warp people's perceptions of partisan groups (Yang et al. 2016).

While online comments present a distorted view of the general public, the counterfactual — a nationally representative public discourse on social media — requires significant effort to observe. Without such a reference point, it is impossible to evaluate the discourse we do observe online or to analyze how it is distorted by selection bias in who participates.

In this paper, we identify the effects of selection into online commenting and algorithmic selection of high-engagement comments on the toxicity of political debate online. First, drawing on multiple data sources, we provide the most comprehensive evidence to date on how commenters differ from non-commenters. Second, we compare real-world Facebook comments scraped from posts published by news organizations to comments elicited on the same posts from a nationally representative sample of Americans using machine learning models built to estimate the toxicity levels of user-generated posts. This design allows us to provide the first estimate of the difference in toxicity between genuine online comments and those written by a nationally representative sample, including people who do not post comments on news articles in the real world.

Furthermore, we consider the role that social media's algorithmic selection plays in amplifying the visibility of toxicity by examining whether toxic comments receive more "likes" than civil ones. Finally, we use an experiment to test whether exposure to toxic comments increases the toxicity of subsequent comments.

Our results indicate that the process of selection into commenting behavior exacerbates the toxicity of online discussion. People who report that they comment frequently on social media not only express more polarized opinions in surveys but write more toxic comments in an elicitation task than other Americans. Similarly, the toxicity of the comments we observe on Facebook substantially exceeds the toxicity of the comments provided on the same articles by a representative sample of the public. Moreover, we find that toxic comments generally attract more "likes" on Facebook than more civil comments. Finally, our survey experiment indicates that people write more toxic comments on a Facebook post when they are randomly exposed to especially toxic comments about the post, suggesting that exposure to toxic comments begets further toxicity. Taken together, our findings suggest that discussions of algorithmic amplification on social media that do not take into account the selection process that we identify will be incomplete.

Prior research on online toxicity

Definitions of online toxicity

Before proceeding, it is first necessary to specify how we define toxic (i.e., uncivil) communication.[1] The specific conceptualizations and operationalizations of incivility vary between studies (Chen 2017; Chen et al. 2019; Coe et al. 2014; Muddiman & Stroud 2017; Mutz 2016; Papacharissi 2004; Rossini 2020; Shmargad et al. 2021; Sobieraj & Berry 2011; Sydnor 2019), but one of the most common themes is the expression of *disrespect* toward others. For example, Sydnor (2019) defines uncivil communication as "any statement that is not respectful of individuals' desire to maintain their self-image" which concurs with Coe et al. (2014)'s definition of incivility: "features of discussion that convey an unnecessarily disrespectful tone toward the

[1]Note: We use "toxicity" and "incivility" interchangeably throughout.

discussion forum, its participants, or its topics."

In this study, we follow Coe et al. (2014), Chen (2017), and Sydnor (2019), who focus on disrespectful tone and word choice in conceptualizing and operationalizing incivility. Specifically, we *define* toxicity (i.e., uncivil) political comments as those expressing disrespect for someone by using insulting language, profanity, or name-calling; by engaging in personal attacks; and/or by employing racist, sexist, and xenophobic terms. As we discuss in detail below, we *measure* toxicity using a machine learning classifier from Google's Perspective API, which is trained to detect "a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion," which is not as detailed as our definition. We validate the Perspective API model by having a subset of our data classified by human coders, who were provided with on our verbatim definition of toxicity and offered additional directives to ensure that their classification of toxic comments reflects our conceptual definition. There was a strong correspondence between the Perspective API score and our estimate based on the coders' classification, which increases our confidence that the Perspective API captured our definition of toxicity (see Appendix E for more information about the validation procedure).

That said, we also note that our arguments and findings pertain to toxicity defined as disrespectful tone and word choice and not necessarily other related constructs such as antidemocratic attitudes or intolerance (Papacharissi 2004; Rossini 2020) that may be expressed without profanity or sexist or xenophobic terms. Future research should consider how different types of toxic discourse emerge on social media and how they might affect people differently, but that is not our focus here.

Implications of online toxicity

Scholars have examined various consequences of exposure to political incivility (Borah 2014; Chen 2017; Gervais 2015; Hwang et al. 2014; Kosmidis & Theocharis 2020; Mutz 2016; Suhay et al. 2018; Theocharis et al. 2016). Overall, they have taken conflicting stances on the normative implications of online incivility for political engagement, deliberation, and polarization. On the one hand, profanity or impolite remarks can be

part of reasoned arguments and serve a useful social purpose in certain contexts among like-minded people (Chen 2017; Chen et al. 2019; Rossini 2020). But on social media such comments may prove toxic to the prospect of conducting a meaningful conversation because seeing one's side being attacked can exacerbate people's animus toward the opposing side (Bail 2021; Settle 2018; Suhay et al. 2018). In this view, civility does not mean simply following arbitrary group norms, but represents a signal of mutual respect, which is a prerequisite for preventing cross-cutting dialogue from spiraling into an exchange of insults (Kingwell 1995; Gervais 2015; Gutmann & Thompson 1996; Cheng et al. 2017; Shmargad et al. 2021; Ziegele et al. 2018).

Despite this ongoing debate, scholars seem to agree that toxicity is an important and useful concept with which to characterize the tenor of political conversations taking place on social media. The purpose of our investigation therefore is to provide clear and comprehensive evidence on whether and how social media amplify toxic discourse.

Evidence of online toxicity

Prior studies have shown consistently that toxicity *exists* online (Ventura et al. 2021; Chen 2017; Coe et al. 2014; Muddiman & Stroud 2017; Sobieraj & Berry 2011; Oz et al. 2018; Papacharissi 2004; Theocharis et al. 2016). However, the severity of online toxicity/incivility seems to vary considerably across different studies depending on the platform examined, the definition of civility used, and the measurement approach (e.g., Chen 2017; Sobieraj & Berry 2011). Importantly, researchers often do not specify a reference point against which to assess how uncivil online public discourse is. This lack of a benchmark appears to have led scholars to draw different normative conclusions from seemingly similar empirical findings. For example, Coe et al. (2014) suggest that the presence of online incivility is "substantial" and "common" after documenting that about one in five comments in a newspaper website were uncivil. Papacharissi (2004) similarly finds that about about 30% of messages in Usenet newsgroups were uncivil or impolite, but concludes, more optimistically, that these features "do not dominate online political dicsussion" and most users expressed their

opinions in a civil manner. In short, studies seeking to measure the extent of online civility has generated different answers from various studies in part because of methodological differences (e.g., different samples) but in part because of a lack of a clearly defined standard or referent.

We advance this literature in the following ways. First, we quantify toxicity from the entirety of Facebook comments ($n = 6,485,910$) on an 11-day census of news articles ($n = 11,305$) published by numerous news sources ($n = 33$). This approach allows us to offer more comprehensive evidence about the prevalence and extent of toxic discourse on social media than most prior research. Further, by eliciting comments from a representative sample of the general public on a subset of the articles, we demonstrate how real-world online discourse among only a relative handful of individuals differs from a counterfactual discourse in which *everyone* discusses the same topics. In this way, we are able to characterize the toxicity of Facebook comments with respect to a clear reference point (what we would observe under full participation) and show how self-selection into comments by a small number of people on Facebook fosters hostile discourse on social media. Finally, we conduct an experiment to investigate the role of algorithmic amplification of top comments, another potential mechanism that could increase toxicity.

Theoretical approach and hypotheses

What makes toxicity prevalent and salient in online public discourse? We hypothesize that the following individual and contextual factors contribute to the emergence and escalation of toxicity on Facebook. First, at the individual level, we expect that people with strong political identities and polarized attitudes are more likely to opt into online commenting (*H1*) and that the over-representation of such individuals amplifies online toxicity (*H2*). At the contextual level, we predict user rating systems on social media (e.g., "liking") favor uncivil comments (*H3*), which may algorithmically amplify toxic discourse. Furthermore, we expect that online toxicity may be further intensified through interpersonal processes where those exposed to others' incivility respond by writing toxic comments themselves (*H4*). Below, we elaborate on each of these

theoretical expectations.

Self-selection of committed partisans into commenting

To understand the sources of online toxicity, it is important to consider who self-selects into posting comments about news articles in the first place. We suggest that political commenting on social media is similar to other forms of political participation in requiring certain levels of ability and motivation. That is, one should have at least some knowledge about a given subject and feel passionately enough about the issue to express a view. Consistent with this argument, online commenters have been found to have stronger partisan attachments, consume more political news, discuss politics more often, turn out to vote at higher rates, and get involved in other political activities more often or with greater verve (Rainie 2012; Settle 2018; Settle et al. 2016; Smith 2013, 2014; Weeks et al. 2017). Building on prior work, we therefore expect that people who self-select into online commenting tend to be more interested in politics, more politically knowledgeable, and have stronger partisan identities.

> *Hypothesis 1a: Frequent online commenters have greater political interest, possess more political knowledge, and identify with a political party more strongly than the general public.*

Another individual factor associated with online political commenting may be *need to evaluate* — the psychological propensity to form strong opinions about various issues (Jarvis & Petty 1996). Prior work shows that those with high need to evaluate — i.e., those who "form opinions about everything" and find it bothersome "to remain neutral" — are more likely to engage in various political activities such as attending a rally and voting (Bizer et al. 2004). There are plausible reasons to expect that this relationship will hold for online commenting as well. Furthermore, since revealing one's political inclinations to others in one's network can be costly (Bode 2017), those who choose to do so would likely hold strong views. Therefore, we consider the following hypothesis:

Hypothesis 1b: Frequent online commenters have higher need to evaluate than the general public.

The expectations that online commenters will have stronger attachments to their party and higher need to evaluate have important implications for how politically polarized they are relative to others. Specifically, partisan strength and political involvement are associated with negative affect towards the opposing party (Iyengar et al. 2012) and adopting extreme policy attitudes (Taber & Lodge 2006). Furthermore, those with higher need to evaluate exhibit stronger political commitment (Federico & Schneider 2007). Another reason why online commenters could be more politically polarized than others may be that frequent exposure to online political communication, in turn, drives people's attitudes and opinions to the extreme (Settle 2018; Suhay et al. 2018).

Building on these studies, and *H1a* and *H1b*, we further hypothesize that those who comment online will have more sharply divided feelings toward the parties (affective polarization) and more polarized opinions by party on major issues (ideological polarization). Prior evidence indicates that Facebook commenting is related to disliking of the out-*party* (Settle 2018), but it remains unclear if this pattern extends to disliking of out-*partisans* or major policy controversies.

Hypothesis 1c: Frequent online commenters have more polarized feelings toward prominent political figures and partisan groups than the general public.

Hypothesis 1d: Frequent online commenters are more widely divided along partisan lines in their ideology and opinions on political issues such as abortion, the Affordable Care Act, and immigration.

Self-selection and online toxicity

To the extent that the participants in online discussion are unrepresentative of the general public, comments that appear on social media may be unrepresentative of the universe of comments that the general public

would have written. In particular, self-selection of committed partisans could plausibly fuel the toxicity of online political discourse.

We expect this pattern because, per Hypotheses 1a–1d, we predict that committed partisans are over-represented in online political discourse (*H1a*) who tend to generate intense opinions about various issues (*H1b*), feel hostile toward political outgroups (*H1c*) and hold polarized political opinions (*H1d*). This group is more likely to express hostile attitudes toward people who do not share their worldview. Another reason is that those who are most vocal on social media are likely the ones who most actively follow partisan news outlets, which often feature uncivil comments and exchanges (Levendusky 2013; Mutz 2016). Furthermore, the presence of hostile people on social media (e.g., trolls) could drive out those unwilling to tolerate incivility (Bor & Petersen 2020; Hmielowski et al. 2014; Sydnor 2019), which, in turn, could make online discourse even more toxic. (While it may seem plausible that this sort of adverse selection effect is stronger online than offline, the evidence remains somewhat unclear (Bor & Petersen 2020)). In short, we expect:

Hypothesis 2a: Online commenters use more toxic language than the general public.

Hypothesis 2b: Real-world online comments are more toxic than the comments generated by the general public.

Algorithmic amplification of toxicity

Importantly, most comments on social media or news websites do not reach many readers, except the few that "go viral" due to organic spread, algorithmic amplification, or the interaction between the two. It is therefore important to consider whether "popular" comments are particularly toxic compared to unpopular ones.

We specifically consider how social media algorithms can make certain comments particularly visible. In other words, we suggest not only that online comments that *exist* on social media are more toxic than how the general public would talk about the same issues, but also that those *seen* by other users tend to be even more toxic due to the way online systems work. There are two theoretical reasons why we expect this

relationship. One possibility is that engagement metrics reward uncivil behavior — more toxic comments will attract more engagement and therefore be more likely to be boosted by algorithms (i.e., go viral). There are two theoretical reasons why we expect this relationship. In addition, humans appear to exhibit a generalized tendency to devote more attention to negative occurrences (Rozin & Royzman 2001), which could increase the likelihood of an emotional response that produces engagement (e.g., Kosmidis & Theocharis 2020). It remains to be seen whether the same pattern holds on social media sites such as Facebook. Therefore, we propose the following hypothesis:

Hypothesis 3: Toxic comments attract more "likes."

Cascade of toxicity from comment to comment

Finally, we expect that algorithmic amplification of high-engagement comments can create a vicious cycle in which toxic comments prompt more toxic replies (e.g., Cheng et al. 2017; Han et al. 2018; Shmargad et al. 2021). There are at least three potential mechanisms of such incivility contagion. First, toxic comments may heighten people's animosity or anger toward the previous commenter or the target attacked by the original comment (Gervais 2015; Settle 2018; Suhay et al. 2018), causing others to respond with hostile remarks themselves (see Mackie et al. 2000). A second potential mechanism is mimicry; people have a tendency to mimic others' language even in online environments (Gonzales et al. 2010; Gervais 2015; Kwon & Gruzd 2017). Third, encountering toxicity can alter people's perceived social norms regarding antisocial behavior, which could provoke them to react in uncivil ways to comply with the norms (Shmargad et al. 2021). In addition to the mechanisms outlined above, these possibilities are consistent with the dynamics of moral contagion documented on social media in which moralized emotional language about political topics is associated with greater diffusion (Brady et al. 2017).[2]

[2]Future research should explore whether emotion, mimicry, or social norms best explain why people post toxic replies. Our data do not allow us to tease out these differences because we did not include measures of potential motivations of writing toxic comments (e.g., anger).

Hypothesis 4a: Toxic comments attract more follow-up comments.

Hypothesis 4b: Toxic comments increase the toxicity of reply comments.

Evidence regarding online toxicity contagion remains unclear. Several survey experiments have found that toxicity can spread from comment to comment in this way (e.g., Cheng et al. 2017; Gervais 2015; Ziegele et al. 2018), whereas other experiments have failed to find similar results (e.g., Han et al. 2018; Molina & Jennings 2018; Rösner et al. 2016). Because these studies typically consider one target issue (e.g., abortion), employ constructed (or hand-picked and edited) comments, and test their effects with unrepresentative samples, it is difficult to know which findings best reflect the general patterns on social media. Some studies find that the toxicity of preceding comments is associated with toxicity of subsequent ones in real world settings, but without randomizing exposure to incivility (Kwon & Gruzd 2017; Shmargad et al. 2021). Our approach offers increased internal *and* external validity due to our use of extensive comments data from Facebook as well as data from a representative survey experiment in which participants were randomly assigned to see actual comments on news articles covering a wide range of issues of the day.

Methods

Data

The data for the analyses presented in this study include news articles and comments scraped from Facebook, an original national public opinion survey, the Pew Research Center's American Trends Panel, and the 2016 American National Election Study (ANES).[3]

We first scraped the posts of news articles from Facebook pages of 33 mainstream news outlets. We were able to gather the content of 11,305 posts and 6,485,910 comments from October 6–16, 2018.[4] Our original

[3]The data and code needed to replicate the results are available on `https://doi.org/10.7910/DVN/SPEOCW`.
[4]For scraping, we used the Netvizz application (Rieder 2013). Our list of news outlets was created based on the Pew Research Center's 2014 report on media trust (Mitchell et al. 2014). See Appendix A for more information.

national survey data was then collected by YouGov from October 31–November 8, 2018 and assembled using a matching and weighting algorithm to approximate a nationally representative sample. The resulting set of 2,200 respondents closely resembles the national population.[5] Our respondents are 76% white, 51% female, and 47% ages 18–44. Approximately one in four graduated from college (29%). Politically, 37% identify as Democrats and 27% as Republicans (45% and 36%, respectively, including leaners).[6]

During the survey, we asked YouGov respondents to comment on articles drawn from Facebook. Each respondent was randomly assigned to see three different articles from the Facebook sample. In each case, respondents were presented with a Facebook post from a news outlet and asked whether they would write a comment if they saw the post on Facebook and subsequently what they would write. (See Appendix B for an example.) The probability that an article was presented to respondents was weighted by total engagement levels (the sum of likes, comments, reactions, and shares it received). This process ensured that the posts shown to YouGov respondents were representative of the news articles people were most likely to see on Facebook. Those who indicated they would not write a comment were asked what they would write if they had to do so. This task was repeated three times per respondent. We compare the resulting comments directly to authentic Facebook comments posted on the same set of articles.

The comment elicitation task also included an embedded survey experiment. To test whether prior comments cause people to make uncivil comments themselves (especially if the prior comments are uncivil), we randomized whether respondents were shown two real comments from Facebook on each post or not.[7] When assigned to the comments condition, respondents were specifically shown the two comments with the most likes for each post, which are frequently highlighted by the Facebook algorithm.

Finally, we also draw on two additional nationally representative surveys: the American Trends Panel

[5]We use survey weights provided by YouGov in our descriptive analyses to best approximate the national population. However, our experimental results are unweighted per Franco et al. (2017) and Miratrix et al. (2018).

[6]The demographic composition of the sample is calculated using survey weights.

[7]The random assignment was conducted at the respondent-comment level. That is, most respondents saw prior comments on some articles but not others.

(ATP), conducted by Pew Research Center, and the 2016 survey of the American National Election Studies (ANES). Pew's ATP surveys have been conducted since 2014. A total of 18,720 respondents have been recruited over the years, although our analyses focus on survey responses from 2016. The ANES 2016 survey was conducted both face-to-face and online with a total of 4,270 respondents.

Measurement

Comment toxicity

To measure comment toxicity, we used the toxicity machine learning model implemented in Google's Perspective API (Wulczyn et al. 2017).[8] This model is trained on labeled data from sources including human moderator-tagged online comments in Wikipedia's talk pages and the *New York Times* comments section. The toxicity score, which ranges from 0–1, indicates the predicted proportion of annotators classifying the comment as "a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion." The Perspective API was found to accurately detect toxicity when compared against human coders' classification of comments on Reddit (Rajadesingan et al. 2020), the *New York Times* (Muddiman et al. 2019), and Facebook and Twitter (Hopp et al. 2020). The API has been used to detect toxicity on various social media platforms (e.g., Hopp et al. 2020; Rajadesingan et al. 2020), including Facebook comments in particular (Ventura et al. 2021; Hopp et al. 2020). This measure serves as the primary dependent variable in our analyses below.

We provide examples of comments at different levels of the toxicity measure in Table B1 in Online Appendix B. Our reading of these examples, as well as other comments in the sample, suggests that the Perspective toxicity score corresponds well — albeit not perfectly — to the textual contents of the comments. In general, comments with low toxicity scores ($< .2$) lack uncivil or derogatory remarks, whereas those with moderate scores (toxicity $\approx .5$) are often quite rude and inflammatory, and those with high scores (toxicity

[8]See `https://developers.perspectiveapi.com/s/about-the-api` for more information about the classifier. The classifier only considers text, not emojis or images.

$>$ .8) tend to be extremely toxic and disrespectful.

To validate the Perspective API more formally in our data, we created two alternative measures of toxicity: using a crowd-sourced pairwise rating approach (Carlson & Montgomery 2017) and creating a dictionary of uncivil words (Muddiman et al. 2019). The procedures and results are detailed in Appendix E. As shown in Figure E1, alternative measures of toxicity — one human-labelled and another dictionary-based — are strongly correlated with the Perspective score.

Commenting behavior

The three national surveys measured people's political commenting behavior using somewhat different survey items. In the ANES survey, respondents were asked whether they have "sent a message on Facebook/Twitter about political issues" in the past 12 months and could respond either "have done" or "have not done." The respondents in the Pew survey were asked how often they "comment, post, or discuss government and politics with others on social media" and could indicate either "often", "sometimes", "hardly ever" or "never." In our YouGov survey, we asked our respondents to indicate how often they write posts on Facebook about content that "concerns politics, public policy, or a controversial social issue such as race, gender, or immigration" on a 9-point scale ranging from "almost constantly" to "never."[9] This variable was collapsed into three categories in the regression models: (1) almost constantly–about once a day; (2) a few times a week–less often; (3) never.

Individual differences and political predispositions

We compare commenters and non-commenters on a variety of political/personal predispositions measured from the three national surveys, which include political interest, political knowledge, partisan strength, need

[9]The item is not intended to capture other activities such as sharing a news story or "liking" others' political posts, but it is possible that some respondents misunderstood the question in that way. Such biases, if present, would lead us to underestimate the differences in political attentiveness and polarization since commenting is more "costly" than other activities such as "liking" (Bode 2017). We thank an anonymous reviewer for making this point.

to evaluate, attitudes toward Trump and Clinton, feelings toward the major parties, feelings toward members of the major parties, ideology, and policy opinions (abortion, the Affordable Care Act, and immigration). Given the large number of variables — many of which are based on multi-item batteries — we relegate detailed information about the measurement of each variable to Appendix B, which provides the full text of survey questions and reliability statistics when applicable. Online Appendix C provides descriptive statistics on all of the variables listed above from the Facebook, YouGov, ANES, and Pew data.

Results

Correlates of commenting

First, who comments? Using the two pre-existing surveys and the original national survey, we investigate how self-reported commenters differ from others in their political attentiveness and attitudes. In Figure 1, we test Hypotheses 1a and 1b by examining whether people who report commenting about political issues online show higher levels of political interest, political knowledge, partisan strength, and need to evaluate.[10]

The results show that commenters' self-reported interest in politics is higher than non-commenters by 12–24 percentage points ($p < .005$ in each case). Habitual commenters are also more likely to be more informed about politics ($p < .005$ in ANES and Pew; $p < .1$ in YouGov) and to identify themselves as strong partisans ($p < .005$ in each case). Furthermore, commenters exhibit significantly stronger need to evaluate ($p < .005$).

In Figure 2, we test Hypotheses 1c and 1d by estimating the extent to which frequent commenters hold more polarized opinions than non-commenters.[11] The figure shows that frequent commenters show greater

[10] As noted earlier, commenting behavior was measured on a binary item in the ANES survey, whereas the Pew and YouGov surveys employ multiple categories. For clarity, infrequent commenters from the latter two surveys are not included in Figure 1. See Online Appendix D for regression estimates in tabular format, including differences in political interest, knowledge, etc. (versus non-commenters) for infrequent commenters who were omitted from the figure.

[11] Pure independents are not included in this analysis. Corresponding regression estimates are provided in Tables D2–D4 in Online Appendix D.

affective polarization across different measures and data sources and that that commenters are more ideologically polarized and express more polarized views on specific issues than do non-commenters.

Comment incivility

We now turn our attention to Hypothesis 2 and consider the effect of self-selection into commenting on the civility of online discourse.

Figure 3 plots the distribution of comment toxicity in our YouGov sample by those who self-report that they comment on Facebook about once a day or more (red dashed line) versus those who do not comment at all (black solid line). We find that frequent Facebook commenters leave slightly more toxic comments than those who never comment about politics on Facebook, which is consistent with Hypothesis 2a.[12]

However, Figure 3 considers only the comments posted in our elicitation task. How do those comments compare to those posted in the real world? Figure 4 tests Hypothesis 2b by directly comparing the distribution of comment toxicity between our YouGov sample and real-world Facebook comments. The solid line represents the distribution of comment toxicity from our nationally representative sample. The bulk of the distribution is at the low end of our toxicity scores, though there is a long tail. By contrast, the dashed line showing the distribution of comment toxicity from Facebook indicates that aggressive comments are much more common there.The average toxicity scores for the Facebook and YouGov samples are respectively 0.330 and 0.186. In other words, toxicity of real world comments surpasses toxicity of comments by the general public by 77$ — a substantial gap. In Figure D3 in Online Appendix D, we additionally show that the difference in toxicity between the Facebook and YouGov samples persists when we drop comments of two words or fewer, suggesting that the difference between YouGov and Facebook comments does not arise because short or low-effort responses (e.g., “no comment”) are more common in the YouGov sample.

[12]The respondents were asked if they would comment on the post if they saw it on social media. In Figure D1 in Online Appendix D, we compare the distributions of toxicity between those who said they would comment with those who said they would not and find more substantial gaps.

It is important to keep in mind that, while the news articles cover a variety of topics, they were all published from October 6–16, 2018. To help establish that our results generalize beyond this limited time frame, in Appendix D we examine toxicity separately for articles whose titles do or do not contain certain keywords related to high-profile current events (the Brett Kavanaugh confirmation hearings and the 2018 midterm elections). We find that Facebook comments are substantially more toxic than those from the YouGov sample for both types of articles, which suggests that the effect of self-selection on toxicity is likely to hold in other contexts.

Although not originally hypothesized, we also explore potential differences across news sources in Appendix D (see Figures D5 and D6). Importantly, Figure D6 demonstrates considerable variation in toxicity across different news sources, which highlights the importance of drawing on a broad spectrum of partisan and non-partisan media when quantifying the severity of online toxicity.[13]

Toxicity and "likes"

We now consider whether more toxic comments attract more likes on Facebook (Hypothesis 3). Figure 5 visualizes how toxicity relates to the number of likes a comment attracts separately for liberal, conservative, and neutral news outlets. Our data partially support Hypothesis 3. Specifically, we find that likes increase with content toxicity up to about the 80$^{\text{th}}$ percentile of the distribution of comments and then declines.[14] The toxicity score that corresponds to the 80$^{\text{th}}$ percentile is 0.53. In other words, the comments that are predicted to be classified as toxic by about 50% of human coders tend to be most popular. As shown in Table B1 in Online Appendix B, the comments in this range of toxicity (0.5–0.6) are fairly inflammatory, even though they contain less profanity than those in the 0.9–1 range. While this pattern is consistent across the ideological spectrum of news outlets, it is stronger for partisan outlets, especially liberal ones.

[13]We thank an anonymous reviewer for suggesting these analyses.
[14]This nonlinear relationship was not originally hypothesized.

Incivility contagion

Finally, we examine whether toxic comments cause other users to write more toxic comments and/or more comments overall in response (Hypotheses 4a and 4b). Figure 6 depicts the relationship between the toxicity of the two most-liked comments of each article — i.e., those most likely to be highlighted as "featured comments" by Facebook's algorithm at the time our study was fielded — and two quantities: the number of total comments that it garners and the toxicity of other comments. Consistent with Hypotheses 4a and 4b, we find that articles with toxic featured comments tend to receive more comments and that when the featured comments are toxic, the content of subsequent comments is more toxic.

However, the Facebook data are observational and cannot distinguish the effects of exposure to toxic top comments from the effects of other attributes of the articles that attracted toxic comments in the first place.

We therefore return to our YouGov survey data to fully test Hypotheses 4a and 4b, estimating the causal effect of exposure to toxic comments. In our survey experiment, we randomly varied whether the two most-liked Facebook comments on the post in question (i.e., the ones most likely to have been actually surfaced to users on the platform) would be displayed.

Figure 7a first illustrates the effects of exposure to other comments on respondents' willingness to offer their own comments by plotting local polynomial fits for both the average toxicity of featured comments and willingness to write comments separately for respondents not shown featured comments (solid line) and those shown featured comments (dashed line).[15] The closely overlapping lines indicate that we find no measurable effect of comment exposure on willingness to comment irrespective of the severity of the toxicity of the comments randomly shown to respondents.

However, Figure 7b shows that exposure to the featured comments increases the toxicity of respondents' elicited comments when the contents of the featured comments themselves are toxic. Exposure to relatively

[15] Participants treated with prior comments could respond either to the article or to a previous comment. By design, those in the control group wrote "top-level" comments.

civil comments has no significant effect on comment toxicity. However, respondents shown very toxic comments write much more toxic comments compared to those shown the same articles without the featured comments. Being exposed to the worst comments in the pool (toxicity $\approx .9$) increases the toxicity of respondents' comments from 0.21 to 0.33, an estimated effect of approximately 0.5 standard deviations when featured comment toxicity is at its maximum. In Table D7 in Appendix D, we verify these graphical findings using linear regression models.

Conclusion

This study examines who comments on social media about politics and the factors associated with uncivil or toxic comments. We make several important contributions. First, we extend previous research on who self-selects into online political commenting (e.g., Settle 2018) using a variety of variables measuring personal and political traits across three national surveys. Building on the conclusions of Settle (2018), Settle et al. (2016), Smith (2013), Smith (2014), and Weeks et al. (2017), we confirm prior research showing that politically involved partisans are most vocal in online public discourse and offer further evidence that commenters hold more intense opinions and polarized attitudes on a range of subjects.

Second, building on Ventura et al. (2021), Coe et al. (2014), Chen (2017), Muddiman & Stroud (2017), Sobieraj & Berry (2011), Oz et al. (2018), and Papacharissi (2004), we provide the most systematic evidence to date on the extent to which online comments express disrespect in a toxic manner. By collecting more than 6 million comments on more than 11,000 news articles from 33 different news sources, we overcome a common limitation in previous research that relies on relatively limited number of comments and/or sources. Indeed, we find sizable differences in comment toxicity across different news sources, which suggests that researchers should take outlet-level variation into consideration when studying incivility of online comments. Furthermore, prior studies have often subjectively characterized the relative prevalence of toxicity or incivility in such samples (e.g., Coe et al. 2014; Papacharissi 2004). We instead compare real-world comments

with those elicited from a nationally representative sample, a new benchmark that allows us to show how unrepresentative online commenters are and the extent to which the toxicity we observe from them differs from what we would observe from the public as a whole. Our analysis shows that Facebook comments are about 77% more toxic than comments by the general public, implying that the distorting prism of social media greatly exaggerates people's hostility toward one another.

Third, we demonstrate that algorithmic selection of top comments can increase the visibility of toxicity in the Facebook ecosystem. Though much research has attempted to quantify the toxicity of comments that *exist* online (e.g., Coe et al. 2014; Sobieraj & Berry 2011), relatively little attention has been paid to whether comments *seen* by other users are particularly toxic. Consistent with Muddiman & Stroud (2017) and Shmargad et al. (2021), we show that more toxic comments generally attract more likes. In this way, user rating systems that amplify highly engaging content could attract further attention to toxic comments, making online discussion as a whole seem even more hostile than it actually is.

Finally, we provide new experimental evidence investigating concerns about the potential effects of online incivility on subsequent discussion (e.g., Gervais 2015). We find no evidence that exposure to the two most-liked comments on an article post — which are frequently featured by Facebook's algorithm — affects people's willingness to comment, but exposure to toxic featured comments does increase the toxicity of the comments respondents write after exposure. Our analysis clarifies mixed prior evidence on incivility contagion (Cheng et al. 2017; Gervais 2015; Kwon & Gruzd 2017; Shmargad et al. 2021; Ziegele et al. 2018) by providing externally valid experimental data showing that toxicity can spread from comment to comment.

Our findings also have important implications for contemporary debates about algorithmic bias and content moderation (Gillespie 2018; Gorwa et al. 2020). In particular, we add to a small but growing body of evidence illustrating how algorithms optimized to boost engagement can contribute to toxic environments on social media. When considered in combination with our evidence on self-selection into commenting, our study raises concerns about the potential for self-reinforcing dynamics in which highly engaged commenters

deter citizens with less solidified opinions from participating in online conversations, which in turn leads to even more toxic and polarizing debates. Our experiment suggests that the practice of featuring the most-engaged comment on a news post can have the potentially unintended effect of encouraging downstream incivility, making insults and name-calling an even more salient feature of online discussion. Social platforms that seek to expand their user base should emphasize features and affordances that minimize disruptive and uncivil behavior that could make them less attractive as participatory spaces. For their part, news organizations committed to fostering healthy debates can implement their own content moderation policies, either via groups that they control or inclusive community-based social norms (Matias 2019). Ultimately, our results call attention to the democratic benefits of expanding participation in online collective debate by designing spaces that reward broad, constructive engagement rather than insults and posturing among a committed few.

We conclude by noting several important limitations of our study. First, our real-world comments were collected exclusively from Facebook. Although it is one of the largest social media platforms, Facebook commenters can differ from those on other platforms such as YouTube, Reddit, and 4chan (e.g., Duggan & Smith 2013).

In addition, the artificiality of the comment elicitation task could have affected our results. On the one hand, the lack of social monitoring might make toxic commenting more likely. However, comments from respondents who say they post comments on Facebook are still generally less toxic than the actual Facebook comments we observe. Alternatively, it is possible that respondents censor themselves due to social desirability concerns or that the online survey environment does not produce the same level of toxicity as a real-world social media feed. Future research should examine how to best approximate real-world commenting behavior in online surveys.

Third, we cannot test over-time dynamics in this study. We have shown that people who are more politically polarized are more likely to engage in toxic commenting behavior and that exposure to toxic comments

can lead users to post increasingly toxic comments of their own, but do not show how or if these effects compound over time. The necessary ingredients for increasing spirals of toxicity — in which toxic comments lead to more toxic comments through increasingly extreme posting behavior and negative selection — are clearly present on Facebook. Future research should examine these spirals and the extent to which they intersect with algorithmic ranking processes (e.g., Ribeiro et al. 2020), which themselves are constantly adapting and evolving (e.g., Munger 2019).

Fourth, we focus on one key attribute of online comments: toxicity. Notwithstanding the growing concern about the incivility of online political discourse, toxicity is not the only dimension in which social media may fail to reflect the general public. Online discussion could amplify extreme views that very few citizens share, or it could center on particular topics in which most people are not interested. Future research should explore how these characteristics (toxicity, extremity, and topic salience) are related to one another, and examine the potential role that self-selection and algorithmic selection play in making each of those attributes a notable feature of political discourse on social media.

On a related note, although we demonstrated that toxicity can fuel further toxicity, we did not examine how it, in turn, affects other aspects of deliberation. Given the findings of prior research that online comments with uncivil remarks often offer reasoned arguments and evidence (Chen 2017; Rossini 2020), it may be possible that toxicity ultimately sparks heated yet constructive exchanges between opposing sides. On the other hand, toxicity could activate in-group favoritism and out-group antagonism, thus motivating people to toe the partisan line instead of engaging with diverse perspectives and new ideas (Bail 2021; Rathje et al. 2021; Suhay et al. 2018). We leave it for future research to adjudicate between these possibilities by assessing the extent to which incivility promotes or impedes political deliberation on social media.

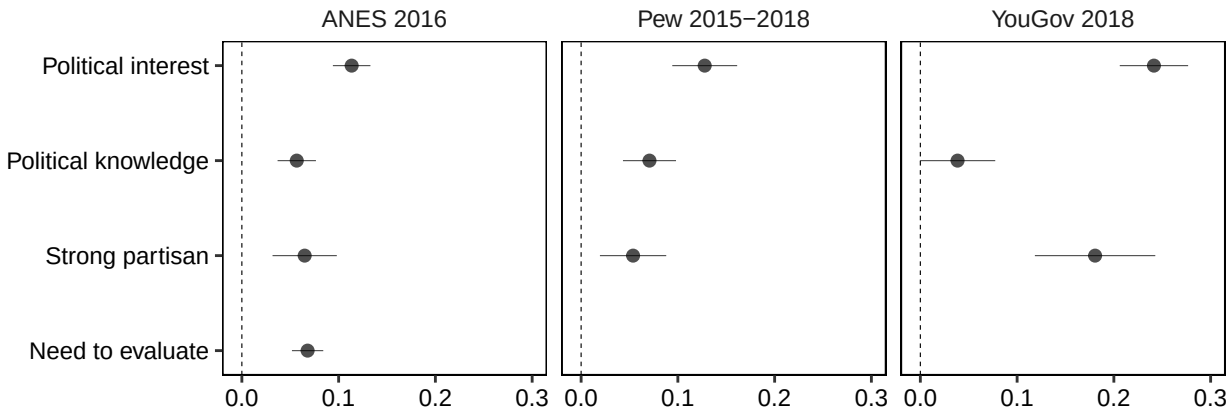
Finally, questions remain about how exposure to uncivil comments influences downstream attitudes and perceptions. The distortions that we document in who comments and which comments are featured could cause people to make false inferences about polarization and toxicity among opposing partisans, for instance.

Similarly, future work should focus more clearly on the mechanisms of how incivility and toxicity affect outcomes.

Despite these limitations, our findings offer an important lesson for research investigating online commenting behavior. Just as the study of voting behavior considers both turnout and vote choice, online commenting behavior consists of both the decision to post and the content of what is written. Research that relies solely on digital trace data to estimate textual features selects on this first outcome, creating potential biases (Nyhan et al. 2017).[16] By eliciting comments from a representative sample and comparing them to what we observe online, we can make valid inferences about both the correlates of posting and the content of the comments that respondents offer. Future research should build on these insights and more carefully consider the role of selection into commenting behavior, exposure to toxicity, and algorithmic amplification in creating toxicity in online comments, an important and highly visible form of incivility in contemporary political life.

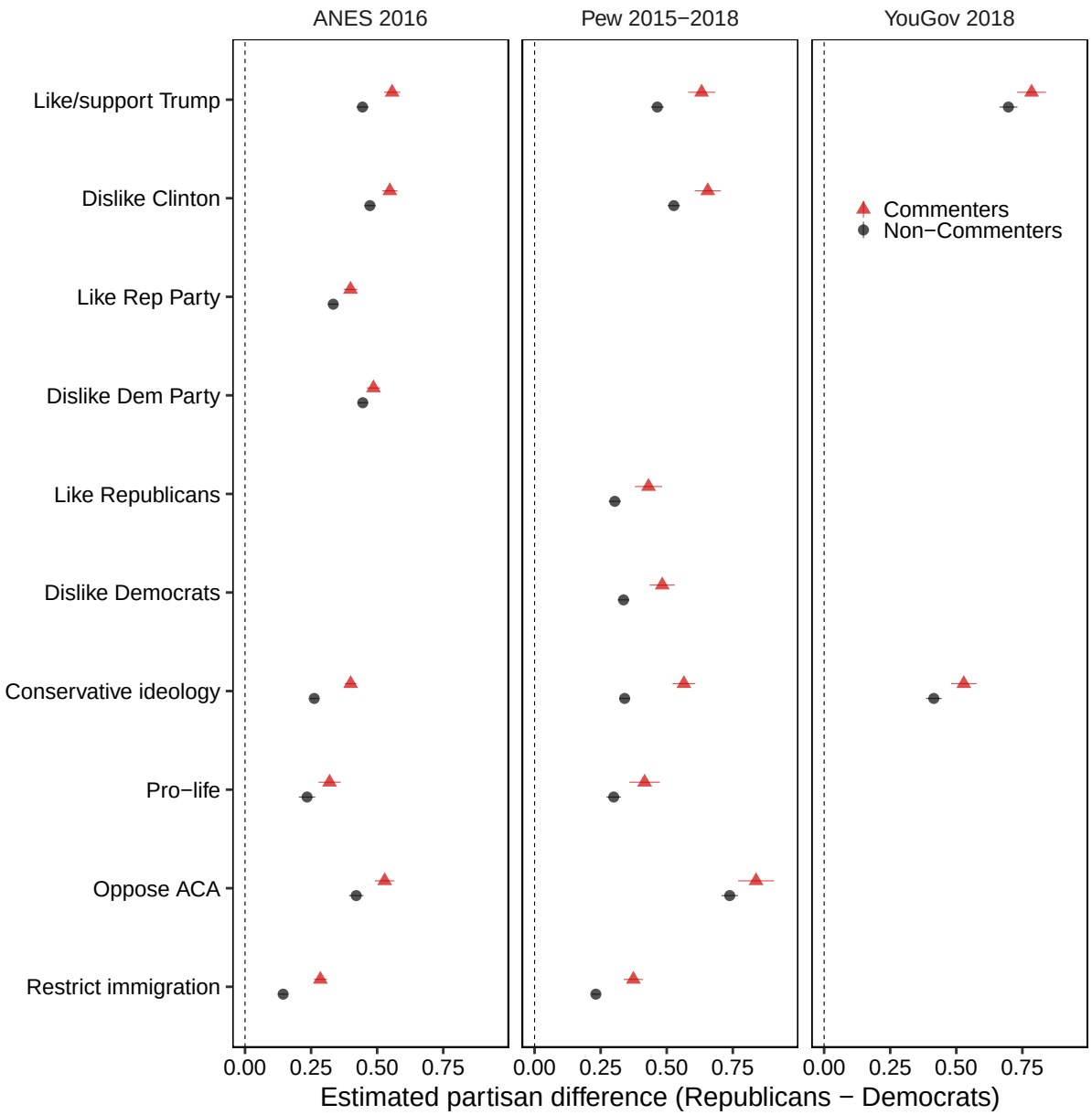
[16]Such biases can be accounted for when digital trace data are linked at the respondent level to representative surveys, though samples with behavioral data from largely closed social platforms such as Facebook are rare and beset by privacy challenges (Guess et al. 2019; Stier et al. 2020).

Figure 1: Differences in political engagement and involvement between online commenters and non-commenters



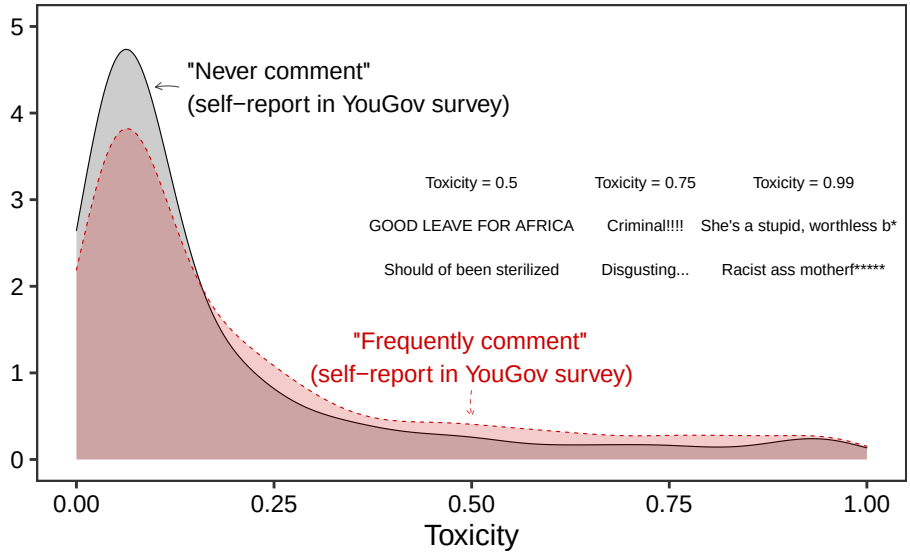
OLS regression estimates with 95% confidence intervals. DVs are the forms of political engagement/involvement. Featured IV is an indicator variable for non-commenters (0) and commenters (1). Circles indicate mean differences between self-reported online commenters and non-commenters. Full regression estimates and information about measures are provided in table D1 in Online Appendix D. All variables are rescaled to 0–1.

Figure 2: Partisan polarization among online commenters and non-commenters



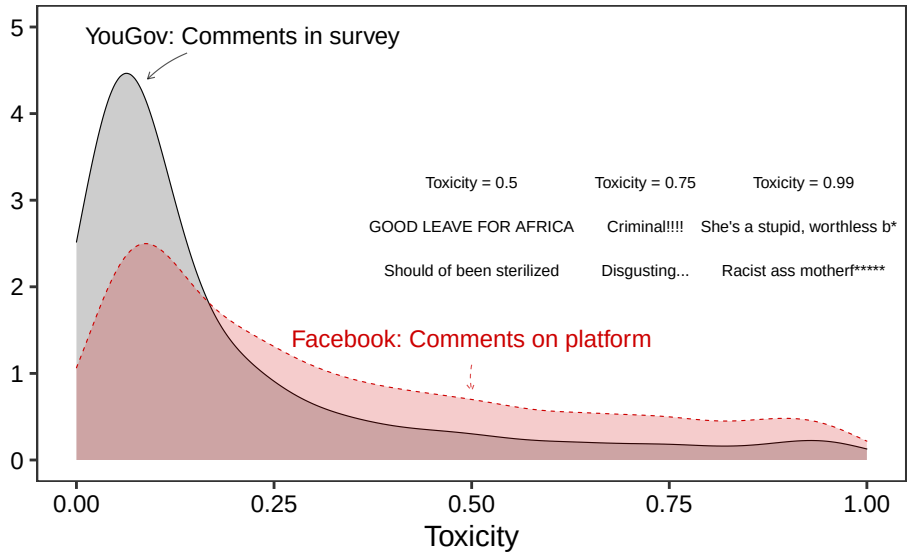
OLS regression estimates with 95% confidence intervals. Triangles indicate the partisan difference (Republicans minus Democrats) among self-reported online commenters. Circles indicate the partisan difference (Republicans minus Democrats) among non-commenters. Full regression estimates are provided in Tables D2–D4 in Online Appendix D. All variables are rescaled to 0–1 where 1 indicates the most pro-Republican attitudes. The partisan gap is significantly wider among commenters than non-commenters in all cases ($p < 0.05$).

Figure 3: Comment toxicity between frequent commenters and others (YouGov)



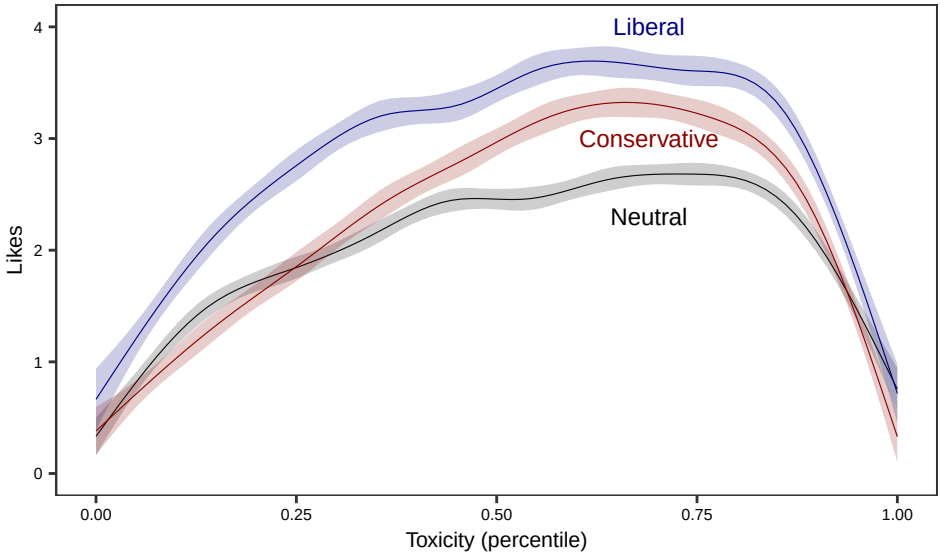
Kernel density estimates of the distribution of the toxicity scores from Google's Perspective API. Mean comment toxicity among non-commenters, infrequent commenters (not shown in this figure), and frequent commenters is 0.18, 0.19 and 0.22, respectively. The mean difference between non-commenters and frequent commenters (0.04) is statistically significant ($p < 0.005$). Table D5 in Online Appendix D provides these estimates in tabular format.

Figure 4: Comment toxicity on Facebook versus a nationally representative sample



Kernel density estimates of the distribution of the toxicity scores from Google's Perspective API. This figure compares 1,188,454 Facebook comments posted on 457 news articles sampled for the YouGov survey instrument with 6,567 comments solicited in the YouGov survey on the same set of articles. Comments on the 10,851 articles not included in the YouGov survey were omitted for this figure. The average comment toxicity was 0.19 for YouGov comments and 0.33 for Facebook comments. The mean difference between non-commenters and frequent commenters (0.14) was statistically significant ($p < 0.005$). Table D6 in Online Appendix D provides these estimates in tabular format.

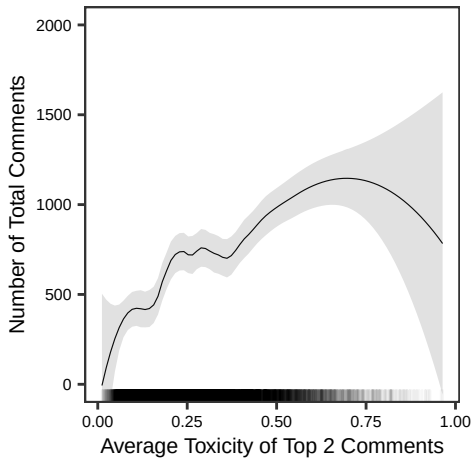
Figure 5: Comment likes and toxicity



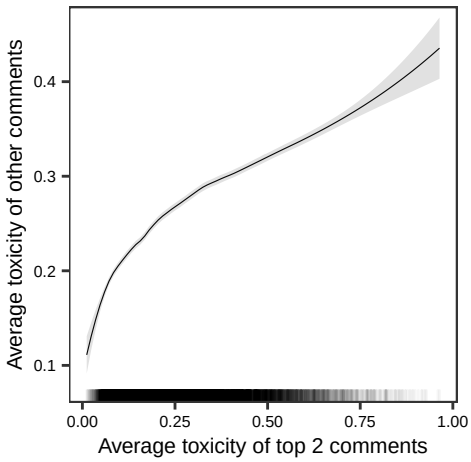
Local polynomial fits of the relationship between comment likes and the percentile of toxicity, separately estimated for neutral, liberal and conservative outlets ($n = 6,485,910$).

Figure 6: Relationship between toxicity of top comments and features of other comments

(a) Number of other comments



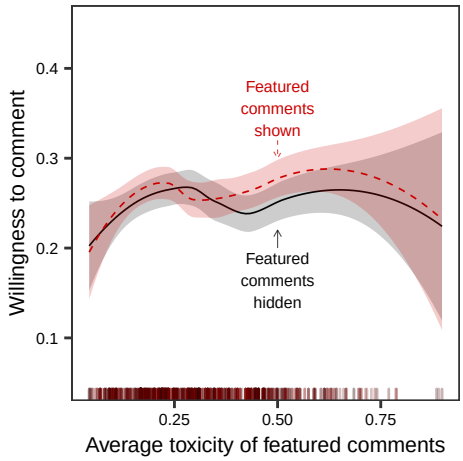
(b) Toxicity of other comments



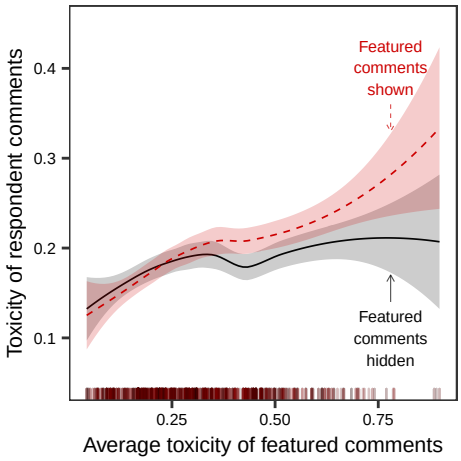
Local polynomial fits estimating the relationships between average toxicity of top 2 comments and (a) number of other comments and (b) toxicity of other comments ($n = 11,252$).

Figure 7: Effects of comment exposure on willingness to comment and toxicity of respondent comments

(a) Willingness to comment



(b) Toxicity of respondent comments



Local polynomial fits estimating the relationship between average toxicity of top 2 comments and (a) willingness to write follow-up comments and (b) toxicity of respondents' comments.

References

Ahler, D. J. (2014). Self-fulfilling misperceptions of public polarization. *The Journal of Politics*, *76*(3), 607–620.

Bail, C. (2021). *Breaking the Social Media Prism*. Princeton University Press.

Bizer, G. Y., Krosnick, J. A., Holbrook, A. L., Christian Wheeler, S., Rucker, D. D., & Petty, R. E. (2004). The impact of personality on cognitive, behavioral, and affective political processes: The effects of need to evaluate. *Journal of personality*, *72*(5), 995–1028.

Bode, L. (2017). Gateway political behaviors: The frequency and consequences of low-cost political engagement on social media. *Social Media+ Society*, *3*(4), 2056305117743349.

Bor, A., & Petersen, M. B. (2020). The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. Unpublished manuscript, June 30, 2020. Downloaded November 18, 2020 from `https://psyarxiv.com/hwb83/`.

Borah, P. (2014). Does it matter where you read the news story? interaction of incivility and news frames in the political blogosphere. *Communication Research*, *41*(6), 809–827.

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.

Carlson, D., & Montgomery, J. M. (2017). A pairwise comparison framework for fast, flexible, and reliable human coding of political texts. *The American Political Science Review*, *111*(4), 835.

Chen, G. M. (2017). *Online incivility and public debate: Nasty talk*. Springer.

Chen, G. M., Muddiman, A., Wilner, T., Pariser, E., & Stroud, N. J. (2019). We should not get rid of incivility online. *Social Media+ Society*, *5*(3), 2056305119862641.

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, (pp. 1217–1230).

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679.

Delli Carpini, M. X., & Keeter, S. (1996). *What Americans know about politics and why it matters*. Yale University Press.

Druckman, J. N., & Levendusky, M. S. (2019). What do we measure when we measure affective polarization? *Public Opinion Quarterly*, *83*(1), 114–122.

Duggan, M., & Smith, A. (2013). 6% of online adults are reddit users. *Pew Internet & American Life Project*, *3*, 1–10.

Federico, C. M., & Schneider, M. C. (2007). Political expertise and the use of ideology: Moderating effects of evaluative motivation. *Public Opinion Quarterly*, *71*(2), 221–252.

Franco, A., Malhotra, N., Simonovits, G., & Zigerell, L. (2017). Developing standards for post-hoc weighting in population-based survey experiments. *Journal of Experimental Political Science*, *4*(2), 161–172.

Gervais, B. T. (2015). Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology & Politics*, *12*(2), 167–185.

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

Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2010). Language style matching as a predictor of social dynamics in small groups. *Communication Research*, *37*(1), 3–19.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 2053951719897945.

Guess, A., Munger, K., Nagler, J., & Tucker, J. (2019). How accurate are survey responses on social media and politics? *Political Communication*, *36*(2), 241–258.

Guess, A. M. (2015). Measure for measure: An experimental test of online political media exposure. *Political Analysis*, *23*(1), 59–75.

Gutmann, A., & Thompson, D. F. (1996). *Democracy and disagreement*. Harvard University Press.

Haenschen, K. (2019). Self-reported versus digitally recorded: Measuring political activity on facebook. *Social Science Computer Review*, *0*(0), 0894439318813586.
URL https://doi.org/10.1177/0894439318813586

Hall, A. (2019). *Who Wants to Run?: How the Devaluing of Political Office Drives Polarization*. Chicago Studies in American Politics. University of Chicago Press.
URL https://books.google.com/books?id=LEeHDwAAQBAJ

Han, S.-H., Brazeal, L. M., & Pennington, N. (2018). Is civility contagious? examining the impact of modeling in online political discussions. *Social Media+ Society*, *4*(3), 2056305118793404.

Hmielowski, J. D., Hutchens, M. J., & Cicchirillo, V. J. (2014). Living in an age of online incivility: Examining the conditional indirect effects of online discussion on political flaming. *Information, Communication & Society*, *17*(10), 1196–1211.

Hopp, T., Vargo, C. J., Dixon, L., & Thain, N. (2020). Correlating self-report and trace data measures of incivility: A proof of concept. *Social Science Computer Review*, *38*(5), 584–599.

Hwang, H., Kim, Y., & Huh, C. U. (2014). Seeing is believing: Effects of uncivil online debate on political polarization and expectations of deliberation. *Journal of Broadcasting & Electronic Media*, *58*(4), 621–633.

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideologya social identity perspective on polarization. *Public opinion quarterly*, *76*(3), 405–431.

Jarvis, W. B. G., & Petty, R. E. (1996). The need to evaluate. *Journal of personality and social psychology*, *70*(1), 172.

Kingwell, M. (1995). *A civil tongue: Justice, dialogue, and the politics of pluralism*. Penn State Press.

Kosmidis, S., & Theocharis, Y. (2020). Can social media incivility induce enthusiasm? evidence from survey experiments. *Public Opinion Quarterly*, *84*(S1), 284–308.

Kwon, K. H., & Gruzd, A. (2017). Is offensive commenting contagious online? examining public vs interpersonal swearing in response to donald trumpâĂŹs youtube campaign videos. *Internet Research*.

Levendusky, M. (2013). *How partisan media polarize America*. University of Chicago Press.

Levendusky, M. S., & Malhotra, N. (2015). (mis) perceptions of partisan polarization in the american public. *Public Opinion Quarterly*, *80*(S1), 378–391.

Mackie, D. M., Devos, T., & Smith, E. R. (2000). Intergroup emotions: Explaining offensive action tendencies in an intergroup context. *Journal of personality and social psychology*, *79*(4), 602.

Mason, L. (2018). *Uncivil Agreement: How Politics Became Our Identity*. University of Chicago Press.

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

Miratrix, L. W., Sekhon, J. S., Theodoridis, A. G., & Campos, L. F. (2018). Worth weighting? how to think about and use weights in survey experiments. *Political Analysis*, *26*(3), 275–291.

Mitchell, A., Gottfried, J., Kiley, J., & Matsa, K. E. (2014). Political polarization & media habits. Pew Research Center, October 21, 2014. Downloaded March 21, 2019 from https://www.pewresearch.org/wp-content/uploads/sites/8/2014/10/Political-Polarization-and-Media-Habits-FINAL-REPORT-7-27-15.pdf.

Molina, R. G., & Jennings, F. J. (2018). The role of civility and metacommunication in facebook discussions. *Communication studies*, *69*(1), 42–66.

Muddiman, A., McGregor, S. C., & Stroud, N. J. (2019). (re) claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries. *Political Communication*, *36*(2), 214–226.

Muddiman, A., & Stroud, N. J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of communication*, *67*(4), 586–609.

Munger, K. (2019). The limited value of non-replicable field experiments in contexts with low temporal validity. *Social Media+ Society*, *5*(3), 2056305119859294.

Mutz, D. (2016). *In-Your-Face Politics: The Consequences of Uncivil Media*. Princeton University Press. URL https://books.google.com/books?id=aX7RjwEACAAJ

Nyhan, B., Skovron, C., & Titiunik, R. (2017). Differential registration bias in voter file data: A sensitivity analysis approach. *American Journal of Political Science*, *61*(3), 744–760.

Oz, M., Zheng, P., & Chen, G. M. (2018). Twitter versus facebook: Comparing incivility, impoliteness, and deliberative attributes. *New media & society*, *20*(9), 3400–3419.

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New media & society*, *6*(2), 259–283.

Pew Research Center (2019). Partisan antipathy: More intense, more personal. October 10, 2019. Downloaded September 2, 2020 from `https://www.pewresearch.org/politics/2019/10/10/the-partisan-landscape-and-views-of-the-parties/`.

Prior, M. (2014). Visual political knowledge: A different road to competence? *The Journal of Politics*, *76*(1), 41–57.

Rainie, L. (2012). Social media and voting. pew research center.

Rajadesingan, A., Resnick, P., & Budak, C. (2020). Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, (pp. 557–568).

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).

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A., & Meira Jr, W. (2020). Auditing radicalization pathways on youtube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, (pp. 131–141).

Rieder, B. (2013). Studying facebook via data extraction: the netvizz application. In *Proceedings of the 5th annual ACM web science conference*, (pp. 346–355). ACM.

Rösner, L., Winter, S., & Krämer, N. C. (2016). Dangerous minds? effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Computers in Human Behavior*, *58*, 461–470.

Rossini, P. (2020). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, (p. 0093650220921314).

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and social psychology review*, *5*(4), 296–320.

Settle, J. (2018). *Frenemies: How Social Media Polarizes America*. Cambridge University Press.

Settle, J. E., Bond, R. M., Coviello, L., Fariss, C. J., Fowler, J. H., & Jones, J. J. (2016). From posting to voting: The effects of political competition on online political engagement. *Political Science Research and Methods*, *4*(2), 361–378.

Shmargad, Y., Coe, K., Kenski, K., & Rains, S. A. (2021). Social norms and the dynamics of online incivility. *Social Science Computer Review*, (p. 0894439320985527).

Smith, A. (2013). Civic engagement in the digital age. *Pew Research Center*, *25*, 307–332.

Smith, A. (2014). Cell phones, social media and campaign 2014.

Sobieraj, S., & Berry, J. M. (2011). From incivility to outrage: Political discourse in blogs, talk radio, and cable news. *Political Communication*, *28*(1), 19–41.

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field.

Suhay, E., Bello-Pardo, E., & Maurer, B. (2018). The polarizing effects of online partisan criticism: Evidence from two experiments. *The International Journal of Press/Politics*, *23*(1), 95–115.

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

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769.

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A., & Parnet, O. (2016). A bad workman blames his tweets: the consequences of citizens' uncivil twitter use when interacting with party candidates. *Journal of communication*, *66*(6), 1007–1031.

Ventura, T., Munger, K., McCabe, K., & Chang, K.-c. (2021). Connective effervescence and streaming chat during political debates. *Journal of Quantitative Description: Digital Media*.

Weeks, B. E., Ardèvol-Abreu, A., & Gil de Zúñiga, H. (2017). Online influence? social media use, opinion leadership, and political persuasion. *International Journal of Public Opinion Research*, *29*(2), 214–239.

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). International World Wide Web Conferences Steering Committee.

Yang, J., Rojas, H., Wojcieszak, M., Aalberg, T., Coen, S., Curran, J., Hayashi, K., Iyengar, S., Jones, P. K., Mazzoleni, G., et al. (2016). Why are "others" so polarized? perceived political polarization and media use in 10 countries. *Journal of Computer-Mediated Communication*, *21*(5), 349–367.

Ziegele, M., Weber, M., Quiring, O., & Breiner, T. (2018). The dynamics of online news discussions: Effects of news articles and reader comments on users' involvement, willingness to participate, and the civility of their contributions. *Information, Communication & Society*, *21*(10), 1419–1435.

Online Appendix A: Full list of Facebook news pages scraped

We chose the following pages based on the list put together by Mitchell et al. (2014). The authors of that report selected the news sources "so as to ask respondents about a range of news media, both in terms of platform and audience size, including some sources with large mass audiences as well as some niche sources. Most of the sources are drawn from those asked about in past Pew Research Center surveys on media consumption." We excluded Al Jazeera America, Google News and BuzzFeed, and replaced *The Colbert Report* and *The Ed Schultz Show* with *Late Show with Stephen Colbert* and *The Rachel Maddow Show*. We were able to gather the content of 11,305 posts and 6,485,910 comments from October 6–16, 2018. Given that we lack individual-level data on Facebook users, we could not screen our data for posts from likely bots.[1]

In order to code the ideological leanings of these outlets, we calculated calculated partisan composition of each outlet's audience – i.e., the share of Republicans between Republicans and Democrats. We then classified outlets with the proportion of Republicans lower than .4 as "liberal," between .4 and .6 as "neutral," and higher than .6 as "conservative" outlets; and classified *The Rachel Maddow Show* and *The Late Show with Stephen Colbert* as liberal sources.

[1] We are not aware of credible estimates of bot prevalence on Facebook. Even if the proportion of posts from inauthentic accounts is non-negligible, their presence should not affect experimental effect estimates.

Table A1: Full list of Facebook news pages

	News outlet	Share of Republicans	Ideology (3 types)	Average toxicity	N
1	ABC	0.45	neutral	0.27	304700
2	BBC	0.30	liberal	0.19	10549
3	Bloomberg	0.52	neutral	0.24	21324
4	Breitbart	0.95	conservative	0.35	992122
5	CBS	0.42	neutral	0.35	213950
6	CNN	0.37	liberal	0.31	853196
7	The Late Show with Stephen Colbert	0.25	liberal	0.26	15366
8	Daily Kos	0.11	liberal	0.37	24462
9	The Daily Show with Trevor Noah	0.18	liberal	0.29	42038
10	The Economist	0.34	liberal	0.23	39209
11	Fox News Channel	0.72	conservative	0.28	1577292
12	Glenn Beck	0.98	conservative	0.29	28629
13	The Guardian	0.25	liberal	0.25	136613
14	Huffington Post	0.33	liberal	0.34	317701
15	Mother Jones Post	0.04	liberal	0.34	24001
16	MSNBC	0.36	liberal	0.32	178990
17	NBC	0.40	liberal	0.32	214070
18	New Drudge	0.91	conservative	0.30	4836
19	The New Yorker	0.17	liberal	0.29	23039
20	NPR	0.26	liberal	0.25	160377
21	The New York Times	0.26	liberal	0.28	209738
22	PBS	0.30	liberal	0.23	35252
23	Politico	0.38	liberal	0.29	94070
24	The Rachel Maddow Show	0.09	liberal	0.36	105689
25	Rush Limbaugh	0.96	conservative	0.28	41720
26	Sean Hannity	0.96	conservative	0.34	97286
27	Slate	0.18	liberal	0.32	22517
28	The Blaze	0.97	conservative	0.31	47017
29	Think Progress	0.03	liberal	0.32	25745
30	USA Today	0.47	neutral	0.29	190707
31	The Washington Post	0.31	liberal	0.30	189020
32	The Wallstreet Journal	0.49	neutral	0.24	46338
33	Yahoo	0.48	neutral	0.28	198347

Online Appendix B: Instruments and measurement

Perspective API Classifier

We calculated toxicity scores for each comment using the toxicity machine learning model from the Google Perspective API that was trained to identify "rude, disrespectful, or unreasonable comment" that would make people "leave a discussion" (see `https://developers.perspectiveapi.com/s/about-the-api` for details). The following table provides randomly selected Facebook comments by toxicity.

Table B1: Randomly selected comments by toxicity

Toxicity	Text
0.1	Pres.Trump is a wizard.He knows and is aware of his enemies and keeps them close .He definitely has his original strategy in handing matters at hand.There is a perfect time and place for everything underneath the sun.MAGA.God bless America and President Donald Trump.
0.1	what would him meeting with emergency responders accomplish?
0.2	Pack your bags and go!
0.2	W was just laying the groundwork. Anyone remember the Patriot Act?
0.3	This is a bad man .
0.3	We lived in London for four years and Scotland six years and my husband was allowed to own a shotgun for game hunting with his customers...It was licensed,required lock,stock and barrels to be stored separately AND the police had unannounced "drop by inspectionsâĂİ by police who did NOT carry firearms!! I fear Europe is a lost cause and it is heartbreaking!!
0.4	Scarlette fever is a rash associated with strep throat. We don't vaccinate against it because it has a bacterial cause you dope, we give antibiotics for it. And, kids still get it :/
0.4	I believe that this ok considering what filth the left is saying
0.5	Its OK let's go all out and vote this errant politician out simple as ABC the judiciary is under siege by trump goons
0.5	Your failure.
0.6	Dem=Nazi
0.6	It doesn't matter if the issue is a big deal or not. You're not going to fix that before election day and maybe never with the Supreme Court being what it is. I can absolutely guarantee that hand wringing and pearl clutching will fix nothing. Dealing with it is not a big deal. It's an annoying deal. It's a bullshit deal. It's something you shouldn't have to screw with but it's hardly insurmountable.
0.7	coverd face troll...typical leftie!
0.7	The biggest con in American history. He lied himself into the WH.
0.8	Lol what a loser school
0.8	You're such a hoser eh!
0.9	this racist man. what a crook.
0.9	Ugly trasvestry we dont care about you either hoe. You only care about yourself and money
1.0	Phil, you are a fucking idiot. Now please go play in traffic and do America a favor.
1.0	Fuck you Trump you're a stinking piece of shite.

Overview of survey measurement approaches

Commenting behavior

As mentioned in the main text, the three national surveys measured people's political commenting behavior using somewhat different survey items. In the ANES survey, respondents were asked whether they have "sent a message on Facebook/Twitter about political issues" in the past 12 months, to which they responded either "have done" or "have not done." The respondents in the Pew survey were asked how often they "comment, post, or discuss government and politics with others on social media," to which they indicated either "often", "sometimes", "hardly ever" or "never." In our YouGov survey, we asked our respondents to indicate how often they write posts on Facebook about content that "concerns politics, public policy, or a controversial social issue such as race, gender, or immigration" on a 9-point scale ranging from "almost constantly" to "never." This variable was collapsed into three categories in the regression models: (1) almost constantly–about once a day; (2) a few times a week–less often; (3) never.

Prior research has shown that survey responses on political media consumption and social media behavior are often inaccurate (e.g., Guess 2015; Guess et al. 2019; Haenschen 2019). The YouGov item was designed to overcome some of the well-known problems in conventional survey questions by incorporating frequency ranges that are granular and clearly defined — e.g., "about once a day" instead of "often" — and including a relatively detailed explanation of what counts as a political issue (Guess et al. 2019). In addition, whereas the ANES and Pew items may capture online political discussion in general, the YouGov item pertains to *writing* political posts on Facebook in particular. Of course, none of these items is without measurement error. But employing multiple data sources with different survey instruments ensures the robustness of our findings. Please see Online Appendix B for full survey question wordings.

Individual differences and political predispositions

We compare commenters and non-commenters on a variety of political/personal predispositions measured from the three national surveys, which include political interest, political knowledge, partisan strength, need to evaluate, attitudes toward Trump and Clinton, feelings toward the major parties, feelings toward members of the major parties, ideology, opinions on abortion, the Affordable Care act, and immigration. Here, we highlight a few important differences in question wording and operationalization across different surveys.

First, political knowledge was measured using standard batteries (e.g., the ANES and our YouGov survey ask for how many years is a United States Senator elected per Delli Carpini & Keeter 1996). The Pew survey tapped into "visual political knowledge" (Prior 2014) by asking factual questions about pictures, maps, or graphs. In addition, whereas ANES and Pew measured people's attitudes toward Trump using a feeling thermometer question with a 100-degree scale, our YouGov survey measured whether people approved or disapproved of Trump's performance in office. Finally, while the ANES survey measured affective polarization by asking about people's feelings toward "the Republican Party" and "the Democratic Party", the Pew survey measured the concept by asking about feelings toward "Republicans" and "Democrats." As Settle (2018, 230) notes, the relationship between commenting and affective polarization may depend on whether people are asked about the parties or the *voters* identifying with each party (see also Druckman & Levendusky 2019). At the same time, these alternative measurement approaches allow us to assess the robustness of the associations.

In what follows, we provide full text of survey questions and reliability statistics when applicable.

YouGov

Survey section

Facebook use: How frequently do you use Facebook?

-Almost constantly
-Several times a day
-About once a day
-A few times a week
-About once a week
-A few times a month
-Once a month
-Less often
-Never

Facebook commenting: [if Facebook use != never] Now we'd like to ask you about content on Facebook that concerns politics, public policy, or a controversial social issue such as race, gender, or immigration.

Thinking about this kind of content, how often do you write posts on Facebook about this kind of content? (This variable was collapsed into three categories in the regression models: (1) almost constantly–about once a day; (2) a few times a week–less often; (3) never. Those who indicated that they don't use Facebook in the previous item were included in the "never" category)

-Almost constantly
-Several times a day
-About once a day
-A few times a week
-About once a week
-A few times a month
-Once a month
-Less often
-Never.

Political interest: Generally, how interested are you in politics?

- Extremely interested
- Very interested
- Somewhat interested
- Not very interested
- Not at all interested

Political knowledge index: The next set of questions helps us learn what types of information are commonly known to the public. Please answer these questions on your own without asking anyone or looking up the answers. Many people don't know the answers to these questions, but we'd be grateful if you would please answer every question even if you're not sure what the right answer is. It is important to us that you do NOT use outside sources like the Internet to search for the correct answer. Will you answer the following questions without help from outside sources? (The number of correct answers the following 5 items was counted and rescaled to 0–1, where 1 indicates being correct on all questions; Cronbach's $\alpha = 0.72$)

-Yes

-No

Political knowledge item 1: For how many years is a United States Senator elected - that is, how many years are there in one full term of office for a U.S. Senator?

-Two years
-Four years
-Six years
-Eight years
-None of these
-Don't know

Political knowledge item 2: How many times can an individual be elected President of the United States under current laws?

-Once
-Twice
-Four times
-Unlimited number of terms
-Don't know

Political knowledge item 3: How many U.S. Senators are there from each state?

-One
-Two
-Four
-Depends on which state
-Don't know

Political knowledge item 4: Who is currently the Prime Minister of the United Kingdom?

-Richard Branson
-Nick Clegg
-David Cameron
-Theresa May
-Margaret Thatcher
-Don't know

Political knowledge item 5: For how many years is a member of the United States House of Representatives elected - that is, how many years are there in one full term of office for a U.S. House member?

-Two years
-Four years
-Six years
-Eight years
-For life

-Don't know

Partisan strength: We recoded the standard 7-point measure of partisan identification provided by YouGov to a 0–1 scale, where 1 indicates "strong Democrat" or "strong Republican" and 0 indicates pure independents.

Republican: We recoded the standard 7-point measure of partisan identification, provided by YouGov into a binary partisanship variable where 0 indicates Democrats or leaning Democrats, and 1 indicates Republicans or leaning Republicans. Pure independents were dropped from this variable.

Trump support: Do you approve or disapprove of the way Donald Trump is handling his job as President? (This variable was rescaled to 0–1, where 1 indicates "strongly approve.")

-Strongly approve
-Somewhat approve
-Somewhat disapprove
-Strongly disapprove

Ideology: We hear a lot of talk these days about liberals and conservatives. Here is a seven-point scale on which the political views that people might hold are arranged from extremely liberal to extremely conservative. Where would you place yourself on this scale, or have you not thought much about this? (This variable was rescaled to 0–1, where 1 indicates "extremely conservative.")

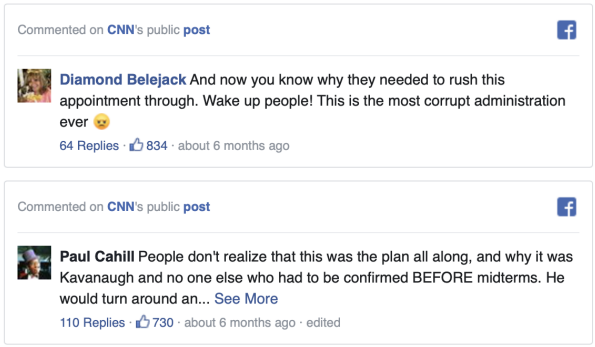
-Extremely liberal
-Liberal
-Slightly liberal
-Moderate; middle of the road
-Slightly conservative
-Conservative
-Extremely conservative
-Haven't thought much about this

Experimental section

The following task was repeated three times per respondent via random draw from a lookup table — i.e., no article preview was shown more than once to a given respondent. Each article was presented with or without the highest-engagement comments it received on Facebook. The randomization was executed at the article level with weights corresponding to total engagement levels (the sum of likes, comments, reactions, and shares that the article received on Facebook).

Please read the Facebook post below.

[sample: article preview]



[sample: featured comments displayed below preview if shown - randomized with probability .5 at commenting task level]

Willingness to write a comment: We are interested in your reaction to the Facebook post above. If you saw it in your Facebook News Feed, would you comment on it? (either on the article or on a previous comment) (This variable was rescaled to 0–1 where 1 indicates "definitely yes.")

-Definitely yes
-Probably yes
-Probably no
-Definitely no

[If yes] What would you write?

[text box]

[If no] If you had to comment on the article or a previous comment, what would you write?

[text box]

American National Election Studies 2016

Commenting behavior: During the past 12 months, have you ever posted a message on Facebook or Twitter about a political issue, or have you never done this in the past 12 months?

- Have done this in past 12 months
- Have not done this in the past 12 months

Partisanship items: Generally speaking, do you usually think of yourself as [a Democrat, a Republican / a Republican, a Democrat], an independent, or what? IF R CONSIDERS SELF A DEMOCRAT: / IF R CONSIDERS SELF A REPUBLICAN : Would you call yourself a strong [Democrat / Republican] or a not very strong [Democrat / Republican]? IF R'S PARTY IDENTIFICATION IS INDEPENDENT, NO PREFERENCE, OTHER, DK: Do you think of yourself as closer to the Republican Party or to the Democratic Party? (We coded these items into the standard 7-point partisanship scale.)

Partisan strength: We recoded the standard 7-point measure of partisan identification, where 1 indicates "strong Democrat" or "strong Republican" and 0 indicates pure independents.

Republican: We recoded the standard 7-point measure of partisan identification, into a binary partisanship variable where 0 indicates Democrats or leaning Democrats, and 1 indicates Republicans or leaning Republicans. Pure independents were dropped from this variable.

Like Trump: Looking at page [PRELOAD: page] of the booklet How would you rate: Donald Trump (This variable was measured on the standard feeling thermometer scale (0–100, where 100 is the warmest feeling), and rescaled to 0–1 where 1 indicates 100.)

Dislike Clinton: Looking at page [PRELOAD: page] of the booklet How would you rate: Hillary Clinton (This variable was measured on the standard feeling thermometer scale (0–100, where 100 is the warmest feeling), and rescaled to 0–1 where 1 indicates 0.)

Like the Republican Party: Looking at page [PRELOAD: page] of the booklet How would you rate: the Republican Party (This variable was measured on the standard feeling thermometer scale (0–100, where 100 is the warmest feeling), and rescaled to 0–1 where 1 indicates 100.)

Dislike the Democratic Party: Looking at page [PRELOAD: page] of the booklet How would you rate: the Democreatic Party (This variable was measured on the standard feeling thermometer scale (0–100, where 100 is the warmest feeling), and rescaled to 0–1 where 1 indicates 0.)

Anti-immigration index: was constructed based on the following 5 items and scaled 0–1 such that 1 indicates the most anti-immigration attitude (Cronbach's $\alpha = 0.83$).

Anti-immigration item 1: Do you think the number of immigrants from foreign countries who are permitted to come to the United States to live should be [increased a lot, increased a little, left the same as it is now, decreased a little, or decreased a lot / decreased a lot, decreased a little, left the same as it is now, increased

a little, or increased a lot]?

Anti-immigration item 2: Now I'd like to ask you about immigration in recent years. How likely is it that recent immigration levels will take jobs away from people already here [extremely likely, very likely, somewhat likely, or not at all likely / not at all likely, somewhat likely, very likely, or extremely likely]?

Anti-immigration item 3: And now thinking specifically about immigrants. (Do you [agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, or disagree strongly /disagree strongly, disagree somewhat, neither agree nor disagree, agree somewhat or agree strongly] with the following statement?) 'Immigrants are generally good for America's economy.'

Anti-immigration item 4: (Do you [agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, or disagree strongly /disagree strongly, disagree somewhat, neither agree nor disagree, agree somewhat or agree strongly] with the following statement?) 'America's culture is generally harmed by immigrants.'

Anti-immigration item 5: (Do you [agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, or disagree strongly /disagree strongly, disagree somewhat, neither agree nor disagree, agree somewhat or agree strongly] with the following statement?) 'Immigrants increase crime rates in the United States.'

Political interest: How interested would you say you are in politics? Are you [very interested, somewhat interested, not very interested, or not at all interested / not at all interest, not very interested, somewhat interested, or very interested]? (This variable was rescaled to 0–1 where 1 indicates "very interested.")

Political interest: (Do you [agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, or disagree strongly /disagree strongly, disagree somewhat, neither agree nor disagree, agree somewhat or agree strongly] with the following statement?) 'Immigrants increase crime rates in the United States.'

Political knowledge index: The number of correct answers the following 4 items was counted and rescaled to 0–1, where 1 indicates being correct on all questions (Cronbach's $\alpha = 0.50$).

Political knowledge item 1: For how many years is a United States Senator elected that is, how many years are there in one full term of office for a U.S. Senator? TYPE THE NUMBER.

Political knowledge item 2: On which of the following does the U.S. federal government currently spend the least? Randomized response option order

- Foreign aid
- Medicare
- National defense
- Social Security

Political knowledge item 3: Do you happen to know which party currently has the most members in the U.S. House of Representatives in Washington?

- Democrats

- Republicans

Political knowledge item 4: Do you happen to know which party currently has the most members in the U.S. Senate?

- Democrats
- Republicans

Need to evaluate index: was constructed based on the following 5 items and scaled 0–1 such that 1 indicates the strongest need to evaluate (Cronbach's $\alpha = 0.74$)..

Need to evaluate item 1: Thinking about yourself, please indicate whether or not the statement is characteristic of you or what you believe: I like to have strong opinions even when I am not personally involved

- Extremely uncharacteristic of me
- Somewhat uncharacteristic of me
- Uncertain
- Somewhat characteristic of me
- Extremely characteristic of me

Need to evaluate item 2: I form opinions about everything

- Extremely uncharacteristic of me
- Somewhat uncharacteristic of me
- Uncertain
- Somewhat characteristic of me
- Extremely characteristic of me

Need to evaluate item 3: It is very important to me to hold strong opinions

- Extremely uncharacteristic of me
- Somewhat uncharacteristic of me
- Uncertain
- Somewhat characteristic of me
- Extremely characteristic of me

Need to evaluate item 4: It bothers me to remain neutral

- Extremely uncharacteristic of me
- Somewhat uncharacteristic of me
- Uncertain
- Somewhat characteristic of me
- Extremely characteristic of me

Need to evaluate item 5: I have many more opinions than the average person

- Extremely uncharacteristic of me

- Somewhat uncharacteristic of me
- Uncertain
- Somewhat characteristic of me
- Extremely characteristic of me

Need to evaluate item 6: I would rather have a strong opinion than no opinion at all

- Extremely uncharacteristic of me
- Somewhat uncharacteristic of me
- Uncertain
- Somewhat characteristic of me
- Extremely characteristic of me

Ideology: Where would you place yourself on this scale, or haven't you thought much about this? (This variable was rescaled to 0–1, where 1 indicates "extremely conservative.")

-Extremely liberal
-Liberal
-Slightly liberal
-Moderate; middle of the road
-Slightly conservative
-Conservative
-Extremely conservative
-Haven't thought much about this

Pro-life: There has been some discussion about abortion during recent years. Which one of the opinions on this page best agrees with your view? You can just tell me the number of the opinion you choose. (This variable was rescaled to 0–1 where 1 indicates the answer that "abortion should never be permitted.")

- By law, abortion should never be permitted.
- By law, only in case of rape, incest, or woman's life in danger.
- By law, for reasons other than rape, incest, or woman's life in danger if need established
- By law, abortion as a matter of personal choice.

Oppose ACA was measured by following items and rescaled to 0–1 where 1 indicates "oppose a great deal":

Do you favor, oppose, or neither favor nor oppose the health care reform law passed in 2010? This law requires all Americans to buy health insurance and requires health insurance companies to accept everyone.

IF R FAVORS THE 2010 HEALTH CARE LAW: Do you favor that [a great deal, moderately, or a little / a little, moderately, or a great deal]?

IF R OPPOSES THE 2010 HEALTH CARE LAW: Do you oppose that [a great deal, moderately, or a little / a little, moderately, or a great deal]?

Pew American Trends Panel Survey instrument

Commenting behavior: Thinking about ways you might use social media sites like Facebook or Twitter. . . How often do you comment, post, or discuss government and politics with others on social media? (This variable was rescaled to 0–1 where 1 indicates "often.")

-Often
-Sometimes
-Hardly ever
-Never

Political interest: How interested are you in keeping up with news and information about the activities of THE FEDERAL GOVERNMENT?

1 Very interested
2 Somewhat interested
3 Not very interested
4 Not at all interested
5 Refused

Political knowledge was measured based on the following 12 items, and rescaled tp 0–1 where 1 indicates being correct on the all 12 answers (Cronbach's $\alpha = 0.71$)..

Political knowledge item 1: What does the line on the map represent?

- Mississippi River
- New Madrid Fault
- Proposed Keystone XL Pipeline
- Expansion of the Midwest Regional Railroad Line

Political knowledge item 2: What is the name of this person?

- Malcolm X
- Martin Luther King, Jr.
- Jesse Jackson
- Thurgood Marshall

Political knowledge item 3: Which country is Pope Francis originally from?

Political knowledge item 4: The United States recently announced that it would re-establish diplomatic relations with which of the following countries?

1 Russia
2 North Korea
3 Cuba
4 Yemen

Political knowledge item 5: Who is Malala Yousafzai?

1 2014 Nobel Peace Prize recipient
2 The first Muslim woman elected to the U.S. Congress
3 An Academy Award-winning director
4 Pakistan's ambassador to the U.S.

Political knowledge item 6: To comply with the health care law, most Americans need to indicate they have health insurance coverage when they...

1 Vote in an election
2 Change their address
3 File their taxes
4 Receive a driver's license

Political knowledge item 7: This graph shows the trend in what national statistic?

1 The inflation rate
2 The corporate tax rate
3 The high school dropout rate
4 The unemployment rate

Political knowledge item 8: Which dot on this map represents where the U.S. military prison at Guantanamo Bay is located?

Political knowledge item 9: There are nine justices on the Supreme Court of the United States. How many are women?

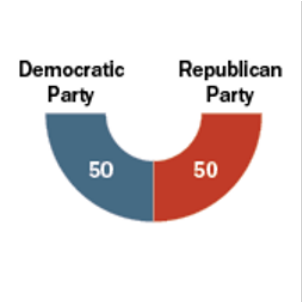
1 One

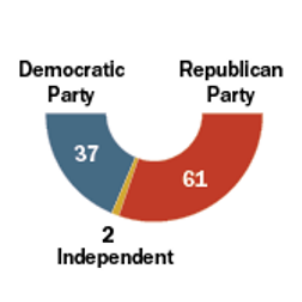
2 Two
3 Three
4 Four

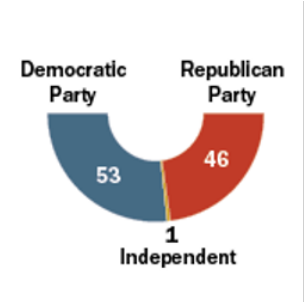
Political knowledge item 10: What country does this person lead?

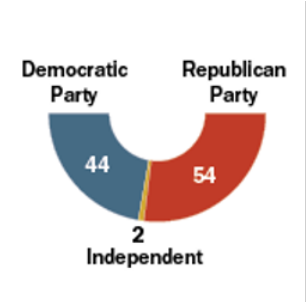
1 North Korea
2 South Korea
3 China
4 Malaysia

Political knowledge item 11: Which of the following shows the number of seats each party holds in the U.S. Senate?









Political knowledge item 12: In what year did the U.S. war in Afghanistan begin? Was it...

1 1997
2 2001
3 2003
4 2010

Partisanship: In politics today, do you consider yourself a [Republican; Democrat; Independent; Something else, please specify] IF INDEPENDENT/SOMETHING ELSE OR MISSING: As of today do you lean more to [The Republican Party; The Democratic Party] (We coded these items into a 5-point partisanship scale.)

Republican: We recoded the 5-point measure of partisan identification, into a binary partisanship variable where 0 indicates Democrats or leaning Democrats, and 1 indicates Republicans or leaning Republicans. Pure independents were dropped from this variable.

Partisan strength: Do you identify with the [Republican/Democratic] Party... [Strongly; Not strongly] (We combined this variable and the 5-point partisanship variable to create a 7-point scale, which was then rescaled to 0–1, where 1 indicates strong partisans, and 1 indicates pure independents. Note that this item was not included in certain waves.)

Like Trump: We'd like to get your feelings toward the presidential candidates on a "feeling thermometer." A rating of zero degrees means you feel as cold and negative as possible. A rating of 100 degrees means you feel as warm and positive as possible. You would rate the group at 50 degrees if you don't feel particularly positive or negative toward the group. How do you feel toward Donald Trump? (This variable was rescaled to 0–1 where 1 indicates 100.)

Dislike Clinton: How do you feel toward Hillary Clinton? (This variable was rescaled to 0–1 where 1 indicates 0.)

Like Republicans: We'd like to get your feelings toward a number of groups on a "feeling thermometer." A rating of zero degrees means you feel as cold and negative as possible. A rating of 100 degrees means you feel as warm and positive as possible. You would rate the group at 50 degrees if you don't feel particularly positive or negative toward the group. How do you feel toward Republicans? (This variable was rescaled to 0–1 where 1 indicates 100.)

Dislike Democrats: How do you feel toward Democrats (This variable was rescaled to 0–1 where 1 indicates 0)

ideology: In general, would you describe your political views as... (This variable was rescaled to 0–1 where 1 indicates "very conservative.")

Very conservative
Conservative
Moderate
Liberal
Very liberal

Anti-immigration index was measured based on the following items, and rescaled to 0–1 where 1 indicates the most anti-immigration attitude (Cronbach's $\alpha = 0.82$).: Thinking about the issue of immigration, how important of a goal should each of the following be for immigration policy in the U.S.? [RANDOMIZE]

-Improving the security of the country's borders
-Increasing deportations of immigrants currently in the country illegally
-Preventing immigrants currently in the country illegally from receiving any government benefits they do not qualify
-Establishing a way for most immigrants currently in the country illegally to stay here legally
-Establishing stricter policies to prevent people who enter the country legally from overstaying their visas and -remaining in the U.S. illegally

-Allowing immigrants who came to the country illegally as children to remain in the U.S. and apply for legal status

Pro-life: was measured using the following variables and rescaled to 0–1 where 1 means the most conservative opinion ('There are no situations at all where abortion should be allowed')

Pro-life item 1: Do you think abortion should be...

- LEGAL in all or most cases;
- ILLEGAL in all or most cases

Pro-life item 2: (ASK IF LEGAL IN ALL/MOST CASES) Which statement comes closer to your own views – even if neither is exactly right?

- There are some situations in which abortion should be restricted

- There are no situations at all where abortion should be restricted

Pro-life item 3: (ASK IF ILLEGAL IN ALL/MOST CASES) Which statement comes closer to your own views – even if neither is exactly right?

- There are some situations in which abortion should be allowed
- There are no situations at all where abortion should be allowed

ACA oppose: Do you approve or disapprove of the health care law passed by Barack Obama and Congress in 2010?

- Approve
- Disapprove

Online Appendix C: Descriptive statistics

Facebook data

Table C1: Post-level descriptive statistics

Variable	Observations	Mean	S.D.	Min	Max	Range	S.E.
Average toxicity of comments	11305	0.26	0.09	0.01	0.93	0.92	0.00
Engagement count	11305	3759.05	9280.91	2	200527	200525	87.29
Comments count	11305	661.13	2098.47	1	70186	70185	19.74

Table C2: Comment-level descriptive statistics

Variable	Observations	Mean	S.D.	Min	Max	Range	S.E.
Toxicity	6485910	0.30	0.26	0	1	1	0.00
Likes count	6485910	2.49	35.62	0	17314	17314	0.01

YouGov data

Table C3: YouGov respondent-level descriptive statistics

Variable	Observations	Mean	S.D.	Min	Max	Range	S.E.
Never comment (base category)	2198	0.45	0.50	0.00	1.00	1.00	0.01
Comment a few times a week or less	2198	0.40	0.49	0.00	1.00	1.00	0.01
Comment once a day or more	2198	0.15	0.36	0.00	1.00	1.00	0.01
Political interest	2199	0.67	0.30	0.00	1.00	1.00	0.01
Political knowledge	2200	0.60	0.32	0.00	1.00	1.00	0.01
Republican (vs. Democrat)	1782	0.44	0.50	0.00	1.00	1.00	0.01
Partisan strength	2108	0.46	0.50	0.00	1.00	1.00	0.01
Support Trump	2200	0.41	0.42	0.00	1.00	1.00	0.01
Conservative	2200	0.50	0.30	0.00	1.00	1.00	0.01

Table C4: YouGov commenting-level descriptive statistics

Variable	Observations	Mean	S.D.	Min	Max	Range	S.E.
Toxicity	6567	0.19	0.22	0.01	0.99	0.99	0.00
Willingness to comment	6598	0.26	0.31	0.00	1.00	1.00	0.00
Average toxicity of featured comments	6586	0.32	0.16	0.04	0.90	0.85	0.00

ANES data

Table C5: ANES descriptive statistics

Variable	Observations	Mean	S.D.	Min	Max	Range	S.E.
Didn't comment in past year (base category)	3646	0.67	0.47	0.00	1.00	1.00	0.01
Commented in past year	3646	0.33	0.47	0.00	1.00	1.00	0.01
Political interest	3639	0.62	0.28	0.00	1.00	1.00	0.00
Political knowledge	4222	0.51	0.30	0.00	1.00	1.00	0.00
Partisan strength	4248	0.38	0.49	0.00	1.00	1.00	0.01
Need to evaluate	3519	0.58	0.23	0.00	1.00	1.00	0.00
Republican (vs. Democrat)	3669	0.47	0.50	0.00	1.00	1.00	0.01
Like Trump	4230	0.37	0.35	0.00	1.00	1.00	0.01
Dislike Clinton	4233	0.58	0.34	0.00	1.00	1.00	0.01
Like the Republican Party	4185	0.44	0.27	0.00	1.00	1.00	0.00
Dislike the Democratic Party	4201	0.52	0.30	0.00	1.00	1.00	0.00
Conservative	3640	0.52	0.24	0.00	1.00	1.00	0.00
Pro-life	4208	0.35	0.37	0.00	1.00	1.00	0.01
Oppose ACA	4264	0.54	0.38	0.00	1.00	1.00	0.01
Strict immigration	3575	0.44	0.23	0.00	1.00	1.00	0.00

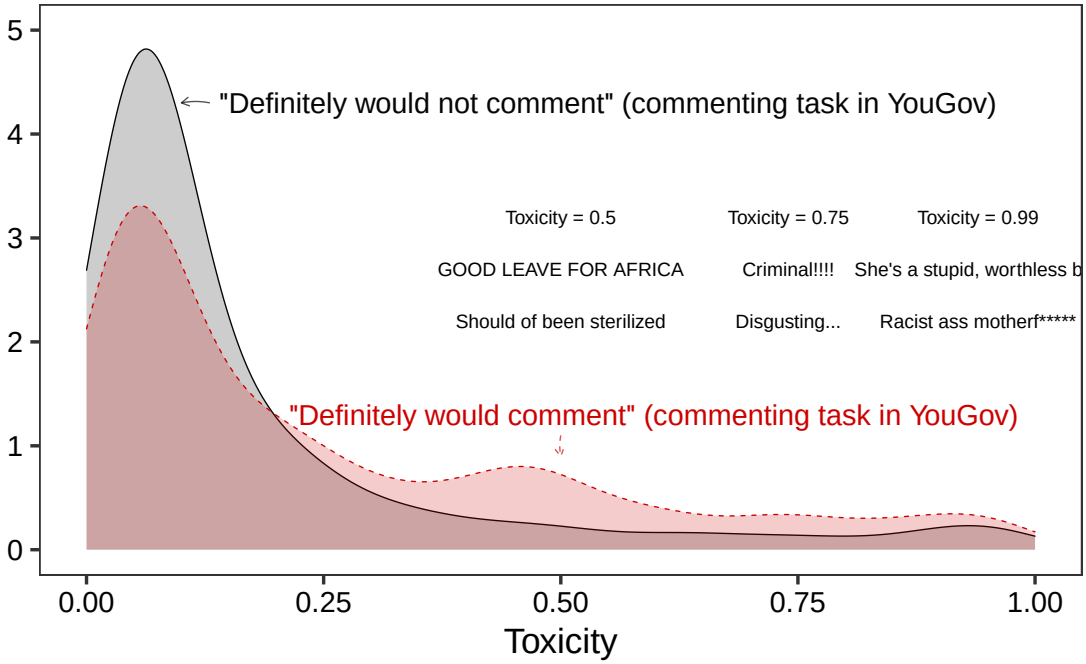
Pew data

Table C6: Pew descriptive statistics

Variable	Observations	Mean	S.D.	Min	Max	Range	S.E.
Never comment (base category)	4563	0.46	0.50	0.00	1.00	1.00	0.01
Comment hardly ever	4563	0.25	0.44	0.00	1.00	1.00	0.01
Comment sometimes	4563	0.20	0.40	0.00	1.00	1.00	0.01
Comment often	4563	0.09	0.29	0.00	1.00	1.00	0.00
Political interest	3205	0.81	0.24	0.00	1.00	1.00	0.00
Political knowledge	3147	0.73	0.20	0.00	1.00	1.00	0.00
Partisan strength	4820	0.69	0.32	0.00	1.00	1.00	0.00
Like Trump	4064	0.32	0.36	0.00	1.00	1.00	0.01
Dislike Clinton	4085	0.58	0.37	0.00	1.00	1.00	0.01
Like Republicans	4661	0.51	0.30	0.00	1.00	1.00	0.00
Dislike Democrats	4667	0.48	0.30	0.00	1.00	1.00	0.00
Conservative	12147	0.49	0.27	0.00	1.00	1.00	0.00
Pro-life	6049	0.45	0.33	0.00	1.00	1.00	0.00
Oppose ACA	3801	0.49	0.50	0.00	1.00	1.00	0.01
Strict immigration	4517	0.60	0.23	0.00	1.00	1.00	0.00

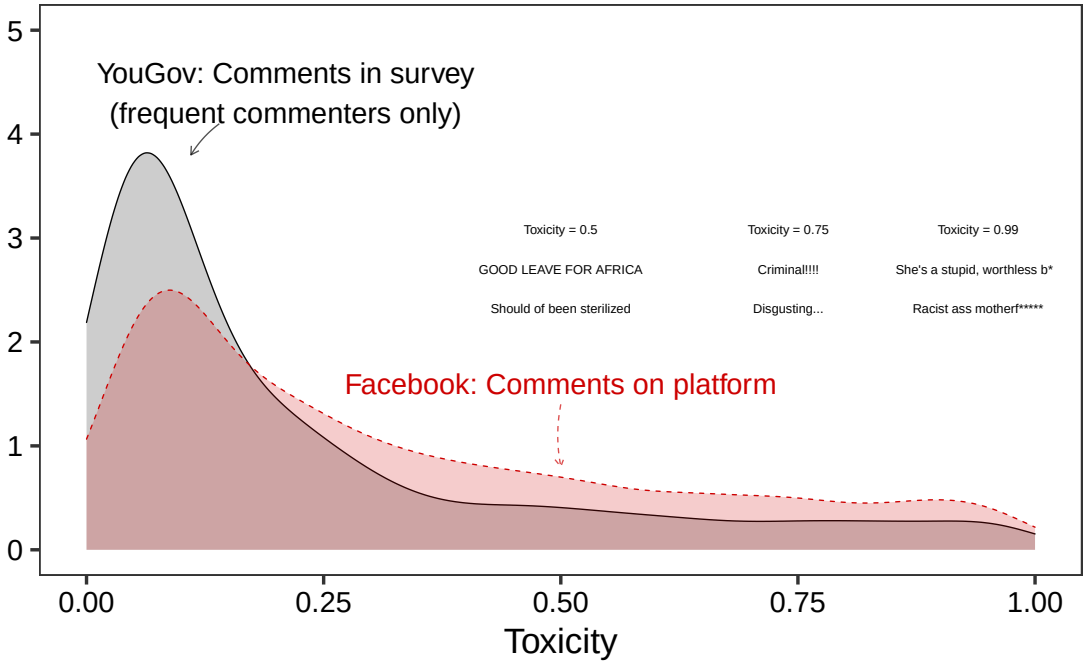
Online Appendix D: Additional results

Figure D1: Comment toxicity by willingness to comment (YouGov)



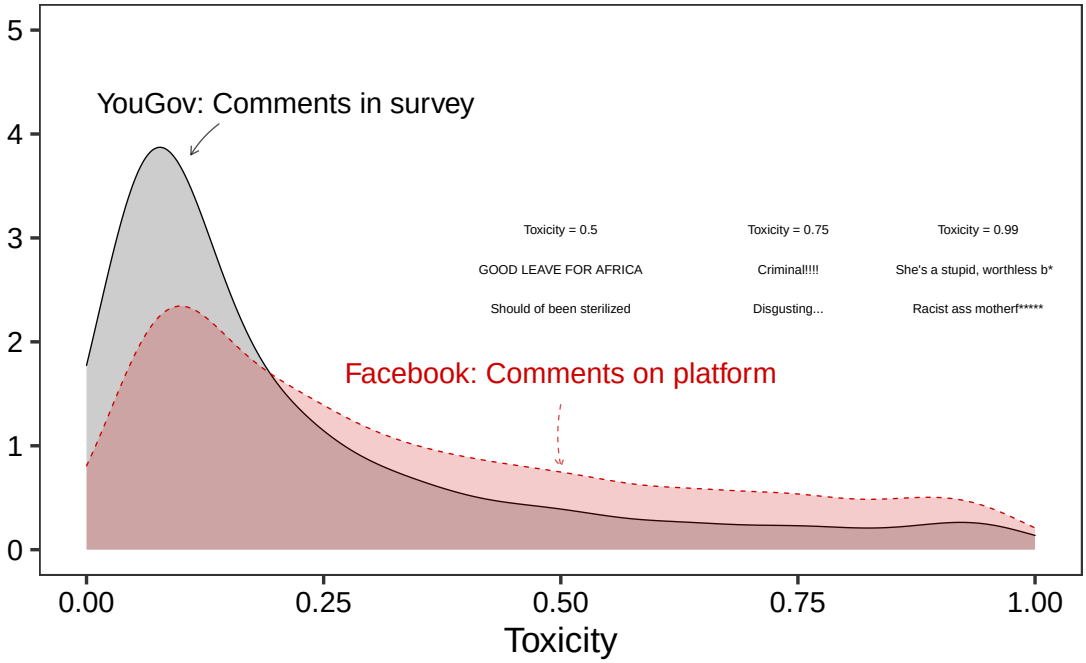
Kernel density estimates of the distribution of the toxicity scores from Google's Perspective API. Mean comment toxicity is 0.17 among those who said they "definitely would not comment," 0.19 among those "probably would not comment" (not shown in this figure), 0.21 among those "probably would comment" (not shown in this figure), and 0.26 among those "definitely would comment" on the presented article. The mean differences between those definitely would not comment and the other groups are all statistically significant ($p < 0.05$ or $p < 0.005$). Table D5 provides these estimates in tabular format.

Figure D2: Comment toxicity on Facebook versus self-reported frequent commenters in a nationally representative sample



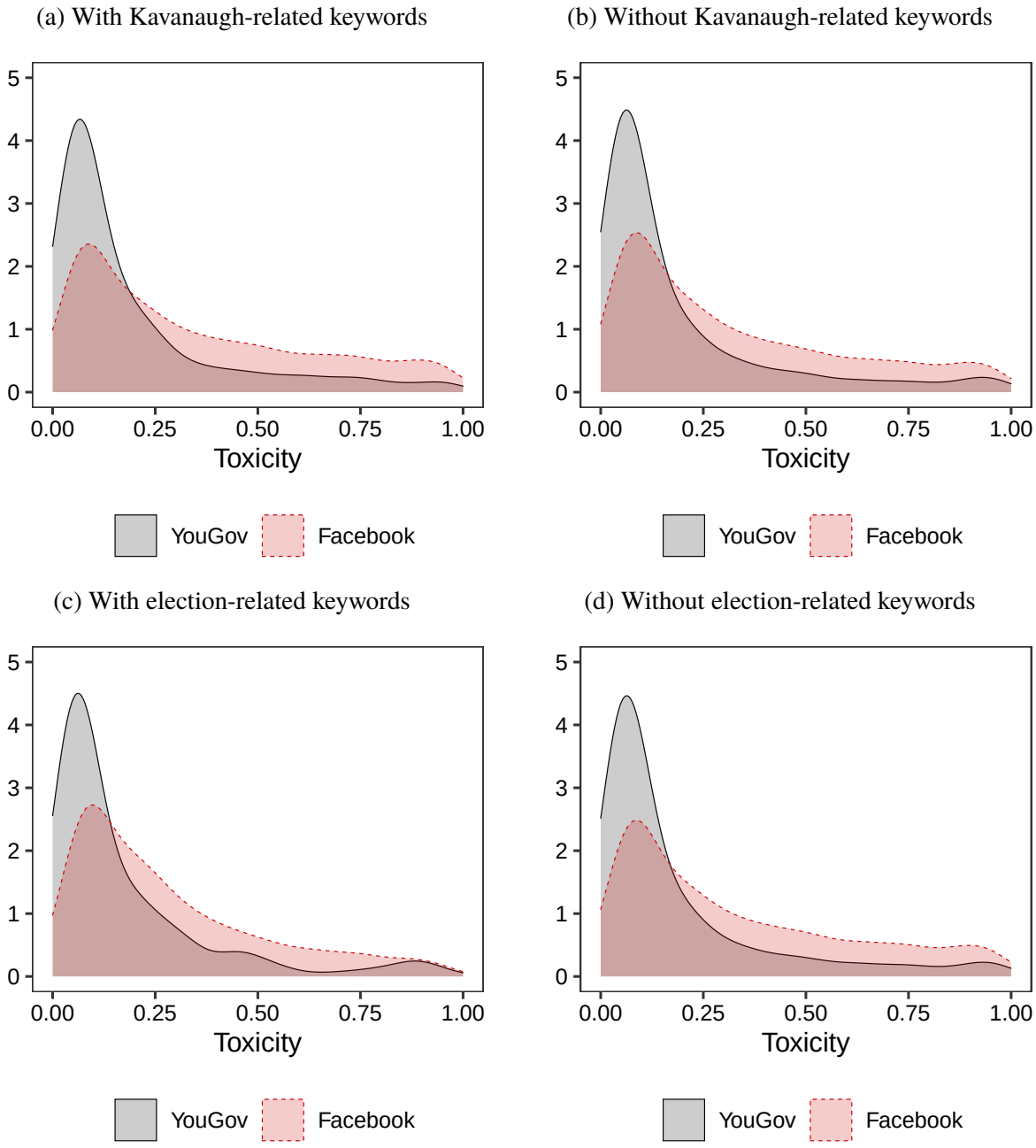
Kernel density estimates of the distribution of the toxicity scores from Google's Perspective API. This figure compares 1,188,454 Facebook comments posted on 456 news articles sampled for the YouGov survey instrument with 1,010 comments solicited in the YouGov survey from self-identified frequent commenters (i.e., those comment about political issues on Facebook once a day or more often). Comments on the 10,851 articles not included in the YouGov survey were omitted for this figure. The average comment toxicity was 0.22 for YouGov comments and 0.33 for Facebook comments. The mean difference between non-commenters and frequent commenters (0.11) was statistically significant ($p < 0.005$). Table D6 provides these estimates in tabular format.

Figure D3: Comment toxicity on Facebook versus a nationally representative sample (comments with two words or fewer dropped)



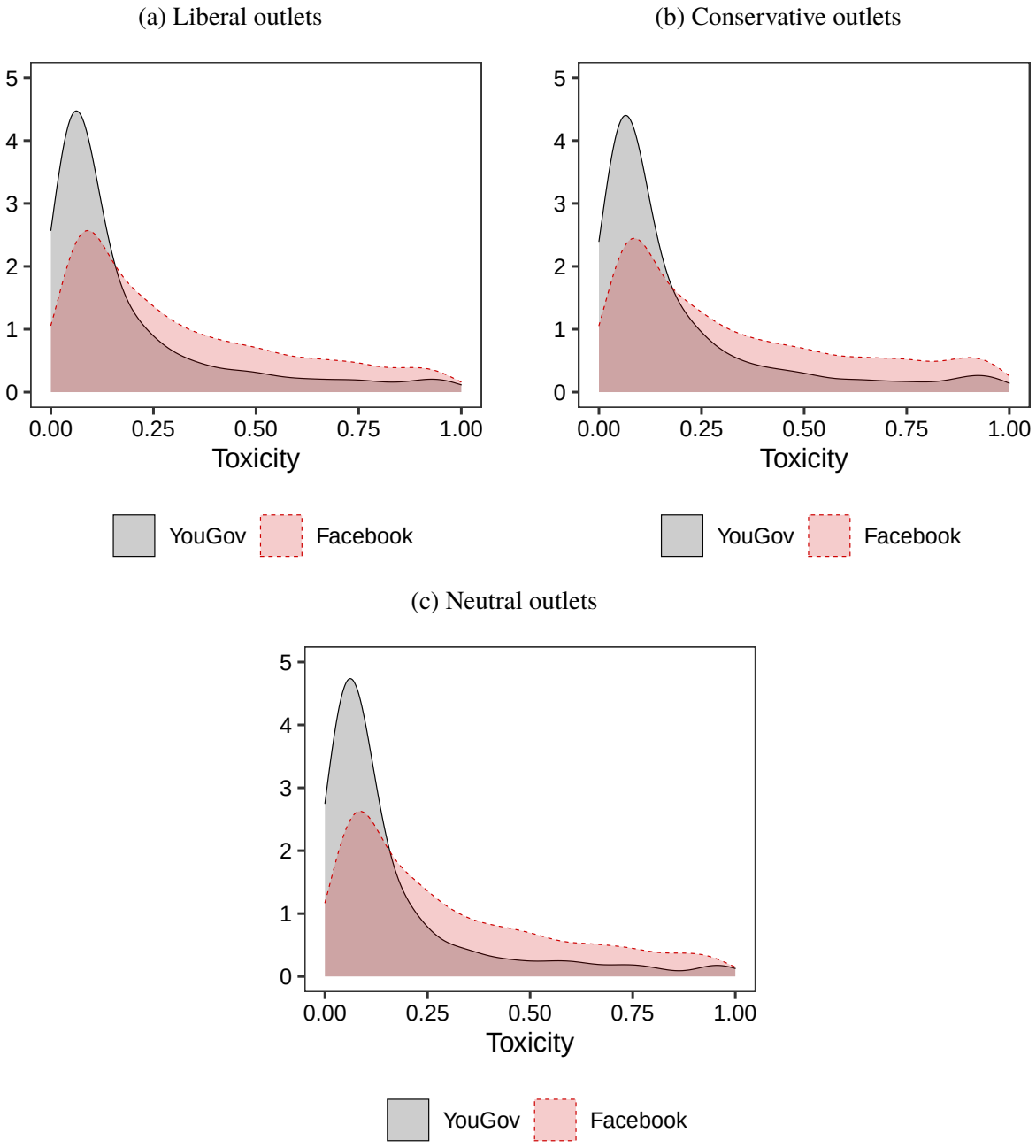
Kernel density estimates of the distribution of the toxicity scores from Google's Perspective API. This figure compares 1,028,427 Facebook comments posted on 456 news articles sampled for the YouGov survey instrument with 4,427 comments solicited in the YouGov survey on the same set of articles. This figure is restricted to comments that consist of 3 or more words. Comments on the 10,851 articles not included in the YouGov survey were omitted for this figure. The average comment toxicity was 0.19 for YouGov comments and 0.33 for Facebook comments. The mean difference between non-commenters and frequent commenters (0.12) was statistically significant ($p < 0.005$).

Figure D4: Comment toxicity on Facebook versus a nationally representative sample by topical keywords



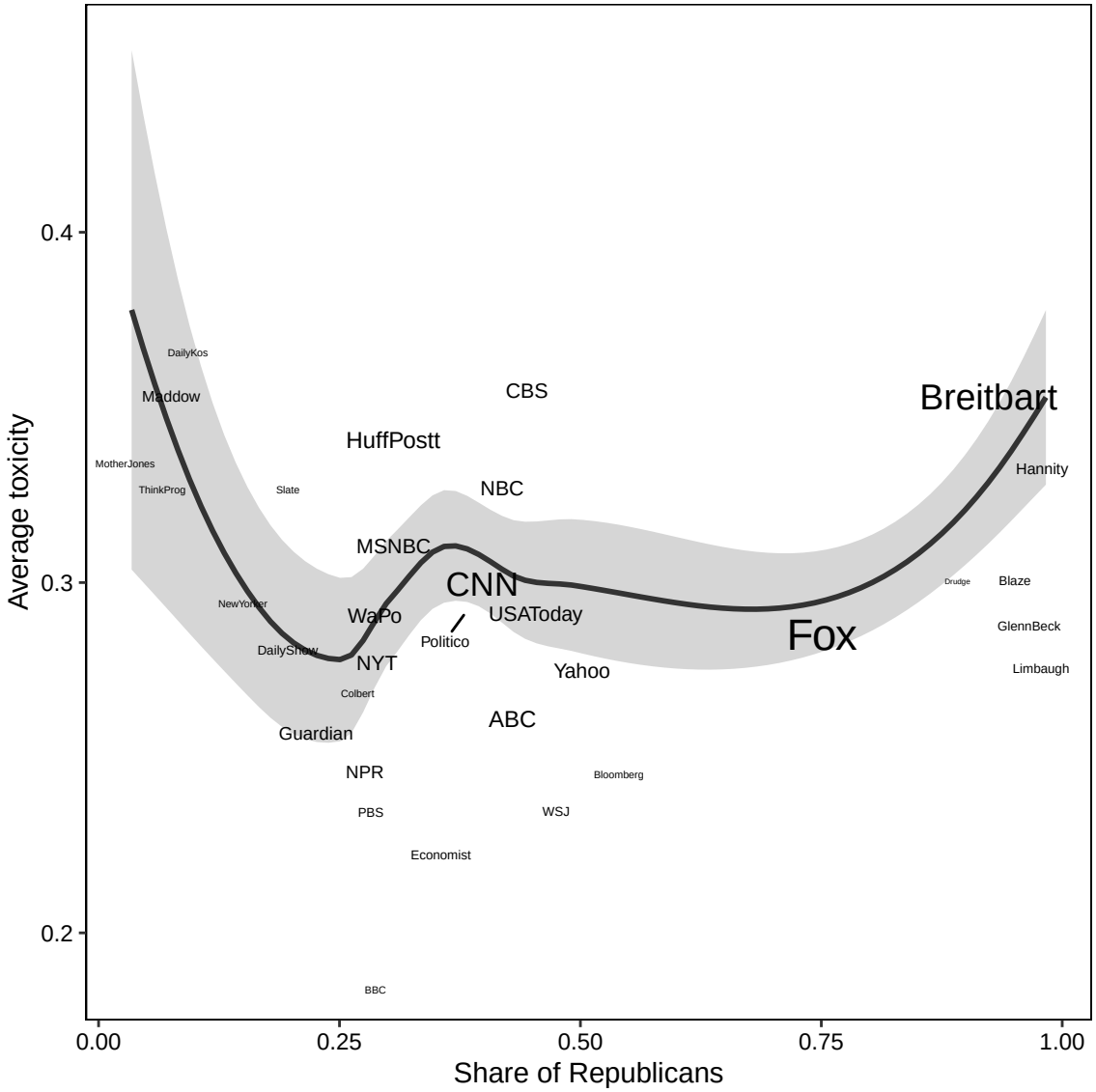
Kernel density estimates of the distribution of the toxicity scores from Google's Perspective API. Panel (a) compares 260,806 Facebook comments posted on articles with titles that include "Kavanaugh," "Ford," "Ramirez," "Swetnick," or "Supreme" with 904 comments solicited in the YouGov survey on the same articles. Panel (a) compares 920,558 Facebook comments posted on articles with titles including none of the Kavnaugh-related keywords with 5,663 comments solicited in the YouGov survey on the same articles. Panel (c) compares 60,547 Facebook comments posted on articles with titles that include "Midterm" or "Election" with 276 comments solicited in the YouGov survey on the same articles. Panel (d) compares 1,120,817 Facebook comments posted on articles with titles including none of the midterm-related keywords with 6,291 comments solicited in the YouGov survey on the same articles. Overall, the four panels show that Facebook comments are substantially more toxic than YouGov comments regardless of whether the article was related to the two major political topics or not, though the difference is slightly less pronounced for articles on the election.

Figure D5: Comment toxicity on Facebook versus a nationally representative sample by partisanship of news sources



Kernel density estimates of the distribution of the toxicity scores from Google's Perspective API. Panel (a) compares 424,328 Facebook comments posted on articles from liberal sources and 3,283 comments solicited in the YouGov survey on the same articles. Panel (b) compares 701,524 Facebook comments posted on articles from conservative sources with 2,685 comments solicited in the YouGov survey on the same articles. Panel (c) compares 599 Facebook comments posted on articles from neutral sources with 62,602 comments solicited in the YouGov survey on the same articles. See Table A1 for the list of news sources and their ideological leanings. Overall, the results show that Facebook comments are substantially more toxic than YouGov comments across news outlets with different ideological leanings, though the pattern is slightly more pronounced in conservative outlets than other types.

Figure D6: Average toxicity of comments by share of Republicans among users of news sources



Local polynomial fit between average toxicity and share of Republicans among users at the news outlet level ($n = 33$), weighted by the number of comments in the Facebook sample (total $n = 6,485,910$). The size of the outlet labels varies by the number of comments on each outlet. Table A1 provides the estimates in tabular form. This figure shows that news sources at the extreme attract comments that are more toxic on average than those around the center. We find no evidence that conservative media attract more toxic comments than liberal media.

Table D1: Regression estimates for Figure 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Data	ANES	ANES	ANES	ANES	Pew	Pew	Pew	YG	YG	YG
DV	interest	knowledge	PID strength	need to evaluate	interest	knowledge	PID strength	interest	knowledge	PID strength
Commented in past year	0.113***	0.057***	0.065***	0.068***						
	(0.010)	(0.010)	(0.017)	(0.008)						
Hardly ever					0.015	0.024**	0.008			
					(0.011)	(0.009)	(0.012)			
Sometimes					0.064***	0.036***	0.039***			
					(0.012)	(0.010)	(0.013)			
Often					0.128***	0.071***	0.054***			
					(0.017)	(0.014)	(0.017)			
A few times a week or less often								0.110***	0.023	0.059*
								(0.013)	(0.015)	(0.024)
About once a day or more often								0.241***	0.038	0.181***
								(0.018)	(0.020)	(0.032)
Constant	0.564***	0.488***	0.352***	0.564***	0.786***	0.730***	0.678***	0.594***	0.583***	0.421***
	(0.006)	(0.006)	(0.010)	(0.005)	(0.007)	(0.005)	(0.007)	(0.009)	(0.010)	(0.016)
Observations	3,636	3,629	3,628	3,516	2,653	2,669	4,343	2,197	2,198	2,106
Adjusted R^2	0.035	0.008	0.004	0.019	0.025	0.011	0.003	0.080	0.001	0.014

* $p < .05$, ** $p < .01$, *** $p < .005$ (two-sided). Cell entries are OLS coefficients with standard errors in parentheses. All dependent variables are rescaled to 0–1, where 1 indicates the most intensive involvement. Figure 2 in the main manuscript displays the coefficients on "Commented in past year" (Columns 1–4), "Often" (Columns 5–7) and "About once a day or more often" (Columns 8–10).

Table D2: Regression estimates for Figure 2 (ANES)

DV	(1) Like Trump	(2) Dislike Clinton	(3) Like Reps	(4) Dislike Dems	(5) Conservative	(6) Pro-life	(7) Oppose ACA	(8) Anti-immigration
Republican	0.444***	0.473***	0.334***	0.445***	0.262***	0.235***	0.421***	0.144***
	(0.012)	(0.011)	(0.010)	(0.009)	(0.008)	(0.016)	(0.014)	(0.009)
Commented past year	−0.060***	−0.018	−0.065***	0.009	−0.100***	−0.082***	−0.054***	−0.091***
	(0.013)	(0.012)	(0.011)	(0.011)	(0.009)	(0.018)	(0.016)	(0.010)
Republican × Commented	0.112***	0.075***	0.065***	0.041*	0.138***	0.085***	0.108***	0.141***
	(0.019)	(0.018)	(0.016)	(0.016)	(0.014)	(0.026)	(0.023)	(0.016)
Constant	0.165***	0.342***	0.289***	0.286***	0.420***	0.275***	0.326***	0.389***
	(0.008)	(0.008)	(0.007)	(0.006)	(0.006)	(0.011)	(0.010)	(0.006)
Observations	3,134	3,133	3,115	3,128	3,151	3,126	3,151	3,103
Adjusted R^2	0.468	0.510	0.405	0.541	0.427	0.129	0.351	0.201

* $p < .05$, ** $p < .01$, *** $p < .005$ (two-sided). Cell entries are OLS coefficients with standard errors in parentheses. All dependent variables are rescaled to 0–1, where 1 indicates the most conservative opinions. The coefficients on Republican is the estimated partisan gaps among non-commenters, and the coefficients on the interaction between Republican and commenting is the difference in the partisan gaps between non-commenters and commenters (i.e., those who did not comment about political issues in the past 12 months). A positive coefficient on the interaction terms indicates the partisan gap is wider among commenters—which was the case in all models. Figure 3 displayed the coefficients on Republican and the sum of the coefficients on Republican and the interaction term.

Table D3: Regression estimates for Figure 2 (Pew)

DV	(1) Like Trump	(2) Dislike Clinton	(3) Like Reps	(4) Dislike Dems	(5) Conservative	(6) Pro-life	(7) Oppose ACA	(8) Anti-immigration
Republican	0.464***	0.527***	0.304***	0.337***	0.341***	0.299***	0.738***	0.232***
	(0.012)	(0.012)	(0.012)	(0.011)	(0.010)	(0.014)	(0.016)	(0.008)
Comment hardly ever	−0.028*	−0.024	0.008	−0.005	−0.022	−0.018	−0.053***	−0.051***
	(0.013)	(0.013)	(0.013)	(0.012)	(0.011)	(0.015)	(0.017)	(0.009)
Comment sometimes	−0.026	−0.032*	−0.036*	−0.018	−0.077***	−0.009	−0.052**	−0.053***
	(0.015)	(0.014)	(0.015)	(0.013)	(0.012)	(0.017)	(0.019)	(0.010)
Comment often	−0.055**	−0.055***	−0.092***	−0.039*	−0.141***	−0.013	−0.057*	−0.088***
	(0.020)	(0.019)	(0.020)	(0.018)	(0.017)	(0.022)	(0.026)	(0.014)
Republican × hardly	0.025	0.019	−0.013	−0.014	−0.003	0.016	0.069**	0.030*
	(0.020)	(0.019)	(0.020)	(0.018)	(0.017)	(0.023)	(0.026)	(0.014)
Republican × sometimes	0.125***	0.081***	0.045*	0.056***	0.106***	0.083***	0.094***	0.078***
	(0.022)	(0.021)	(0.022)	(0.020)	(0.018)	(0.025)	(0.029)	(0.015)
Republican × often	0.168***	0.129***	0.127***	0.146***	0.224***	0.117***	0.099**	0.142***
	(0.029)	(0.028)	(0.029)	(0.027)	(0.024)	(0.033)	(0.038)	(0.020)
Constant	0.105***	0.346***	0.376***	0.323***	0.368***	0.286***	0.166***	0.512***
	(0.009)	(0.008)	(0.008)	(0.008)	(0.007)	(0.009)	(0.011)	(0.006)
Observations	3,805	3,823	4,089	4,096	3,523	3,574	3,450	4,177
Adjusted R^2	0.514	0.574	0.286	0.368	0.484	0.272	0.612	0.363

* $p < .05$, ** $p < .01$, *** $p < .005$ (two-sided). Cell entries are OLS coefficients with standard errors in parentheses. All dependent variables are rescaled to 0–1, where 1 indicates the most conservative opinions. The dependent variables were regressed on Republican (binary) and three dummy variables indicating commenting frequencies ("never" served as the reference category). The coefficients on Republican is the estimated partisan gaps among non-commenters, and the coefficients on the interactions between Republican and commenting dummies are the differences in the partisan gaps between non-commenters and commenters. A positive coefficient on the interaction term between Republican and "often" indicates the partisan gap is wider among those who often comment about political issues on social media—which was the case in all models. Figure 3 in the main manuscript displays the coefficients on Republican and the sums of the coefficients on Republican and the interaction terms between Republican and "often," which are the estimated partisan gaps among those who comment often.

Table D4: Regression estimates for Figure 2 (YouGov)

DV	(1) Support Trump	(2) Conservative
Republican	0.697***	0.415***
	(0.017)	(0.015)
Comment a few times a week or less often on Facebook	0.019	−0.035*
	(0.017)	(0.015)
Comment about once a day or more often on Facebook	0.047*	−0.088***
	(0.022)	(0.019)
Republican × less often	0.042	0.077***
	(0.025)	(0.022)
Republican × more often	0.088**	0.113***
	(0.033)	(0.029)
Constant	0.081***	0.325***
	(0.012)	(0.010)
Observations	1,780	1,780
Adjusted R^2	0.697	0.550

* $p < .05$, ** $p < .01$, *** $p < .005$ (two-sided). Cell entries are OLS coefficients with standard errors in parentheses. All dependent variables are rescaled to 0–1, where 1 indicates the most conservative opinions. The dependent variables were regressed on Republican (binary) and two dummy variables indicating Facebook commenting frequencies ("never" served as the reference category). The coefficients on Republican is the estimated partisan gaps among non-commenters, and the coefficients on the interactions between Republican and commenting dummies are the differences in the partisan gaps between non-commenters and commenters. A positive coefficient on the interaction term between Republican and "more often" indicates the partisan gap is wider among those who comment about political issues on Facebook about once a day ore more often—which was the case in both models. Figure 3 in the main manuscript displays the coefficients on Republican and the sums of the coefficients on Republican and the interaction terms between Republican and "more often," which are the estimated partisan gaps among those who comment once a day ore more frequently.

Table D5: Toxicity of comments by commenters vs. non-commenters (YouGov)

	(1)	(2)	(3)	(4)
Comment a few times a week or less often on Facebook	0.010	0.010		
	(0.007)	(0.007)		
Comment about once a day or more often on Facebook	0.044***	0.042***		
	(0.010)	(0.010)		
Probably would not comment on the presented article			0.015*	0.014*
			(0.007)	(0.007)
Probably would comment on the presented article			0.039***	0.033***
			(0.010)	(0.010)
Definitely would comment on the presented article			0.088***	0.080***
			(0.016)	(0.015)
Constant	0.175***		0.170***	
	(0.005)		(0.005)	
Article fixed effects	No	Yes	No	Yes
Observations	6,561	6,561	6,567	6,567
Adjusted R^2	0.004	0.056	0.011	0.061

* $p < .05$, ** $p < .01$, *** $p < .005$ (two-sided). Cell entries are OLS coefficients with standard errors clustered at the respondent level in parentheses. The dependent variable is the toxicity score generated by Google's Perspective API that ranges between 0 and 1. Habitual political commenting behavior was used as the explanatory variable in Columns 1 and 2. The willingness to comment on the presented article during the commenting task was used as the explanatory variable in Columns 3 and 4.

Table D6: Comment toxicity on Facebook versus a nationally representative sample

	(1)	(2)	(3)	(4)	(5)	(6)
Facebook data	0.143***	0.107***	0.123***	0.087***	0.110***	0.073***
	(0.003)	(0.004)	(0.004)	(0.004)	(0.009)	(0.009)
Constant	0.186***		0.223***		0.220***	
	(0.003)		(0.004)		(0.009)	
Article fixed effects	No	Yes	No	Yes	No	Yes
Less than 2 words dropped	No	No	Yes	Yes	No	No
Frequent commenters only (YouGov)	No	No	No	No	Yes	Yes
Observations	1,195,021	1,195,021	1,032,854	1,032,854	1,189,464	1,189,464
Adjusted R^2	0.001	0.104	0.001	0.088	0.0001	0.103

* $p < .05$, ** $p < .01$, *** $p < .005$ (two-sided). Cell entries are OLS coefficients with standard errors clustered at the respondent level in parentheses. The dependent variable is the toxicity score generated by Google's Perspective API that ranges between 0 and 1. All comments were used in Columns 1 and 2. Comments with two words or fewer were omitted in Columns 3 and 4. In Columns 5 and 6, all YouGov comments were from self-reported frequent commenters (i.e., those who comment about political issues on Facebook once a day or more often).

Table D7: Effects of comment exposure on toxicity

	DV = Willingness to write comment					DV = Toxicity of YG respondent comments				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treatment (top comments shown)	0.011		0.002	0.013	0.004	0.010		−0.018	0.007	−0.019
	(0.008)		(0.017)	(0.008)	(0.017)	(0.006)		(0.012)	(0.006)	(0.012)
Average toxicity of top comments		0.050*	0.036				0.128***	0.087***		
		(0.023)	(0.033)				(0.018)	(0.024)		
Treatment × top comments toxicity			0.027		0.027			0.085*		0.081*
			(0.048)		(0.049)			(0.036)		(0.038)
Constant	0.253***	0.243***	0.242***			0.181***	0.146***	0.154***		
	(0.007)	(0.009)	(0.012)			(0.004)	(0.006)	(0.008)		
Article fixed effects	No	No	No	Yes	Yes	No	No	No	Yes	Yes
Observations	6,598	6,584	6,584	6,598	6,584	6,567	6,553	6,553	6,567	6,553

* $p < .05$, ** $p < .01$, *** $p < .005$ (two-sided). Cell entries are OLS coefficients with robust standard errors in parentheses clustered by respondent. Dependent variable in columns 1 to 5 is a four-point scale of willingness to comment, rescaled to 0-1 such that 0 is "Definitely no" and 1 is "Definitely yes." The dependent variable in columns 6 to 10 is Perspective API's toxicity score. Constant and coefficient on average toxicity of top comments in columns 4, 5, 9, and 10 are subsumed by article fixed effects. In this table, we verify these graphical findings presented in Figure using linear regression models in which we regress the outcome variables on the treatment, the toxicity of the featured comments, and the interaction between the two. As shown in models 1–5, we do not find a significant treatment effect on willingness to comment. Furthermore, we do not find that (randomized) exposure to top featured comments increases toxicity when averaged across all articles (model 6). However, the toxicity of top comments is significantly associated with the toxicity of respondents' comments ($p < 0.005$; model 7). Importantly, this association is much stronger for those assigned to see the featured comments (0.18) than those who weren't (0.09), and the difference between those quantities (0.09) is statistically significant ($p < 0.05$; model 8). These results hold when we include article fixed effects (Columns 9–10).

Online Appendix E: Validation of the Perspective API

To validate the Perspective API in our data, we created two alternative measures of toxicity.

First, we use a crowdsourced pairwise comparison approach (Carlson & Montgomery 2017). To do so, we first stratified each comment dataset (Facebook and YouGov) into 10 groups based on the estimated level of toxicity and randomly sampled 100 comments from each stratum, resulting in a sample of 1,000 Facebook comments and 1,000 YouGov comments.[1] We then randomly formed 20,000 unique pairs of comments so that each comment was compared with 20 other comments. We supplied these paired comments to online workers recruited from Amazon's Mechanical Turk, who were asked to choose the more uncivil comment between the two.

Workers were provided with a detailed codebook with our definition of toxicity — i.e., expressing disrespect for someone using insulting language, profanity, or name-calling; engaging in personal attacks; and/or employing racist, sexist, and xenophobic terms — along with specific directives and examples designed to make sure that their classification reflects our conceptualization of toxicity. Specifically, they were instructed to account for the severity of profanity, the intent behind the use of profanity (disrespect vs. excitement), and the seriousness of accusations. Workers were required to pass a classification test to ensure that they fully understood the instructions and only those workers who classified at least 6 out of 7 comments accurately were eligible to take up the tasks. (The codebook and test are provided in Online Appendix B.) [2] Subsequently, we calculated a pairwise toxicity rating of a comment as the probability that it is classified to be more toxic in 20 comparisons completed by MTurk workers. As shown in Figure E1a, the pairwise rating is strongly correlated with the Perspective score at $r = 0.79$ (left). Moreover, the pattern holds almost identically for both Facebook and YouGov samples (right).

To further ensure that the Perspective API detects toxicity in our data, we developed a dictionary of uncivil words (Muddiman et al. 2019). We first created a list of the 5,000 most frequent stem words in our sample — excluding stop words such as "he" and "the" — and selected 655 features that may have been used as part of derogatory remarks, from which we narrowed down to 295 features that are either profanity (e.g., "f**k", "b***h") or clearly name-calling (e.g., "sicko", "monster").[3] We plot the binary variable indicating whether a comment includes any of the features against the Perspective API in Figure E1b, which shows a nearly perfect correspondence between the API score and the probability that a comment includes one of the uncivil features. Taken as a whole, the Perspective API appears to perform well at detecting toxic comments in our data.

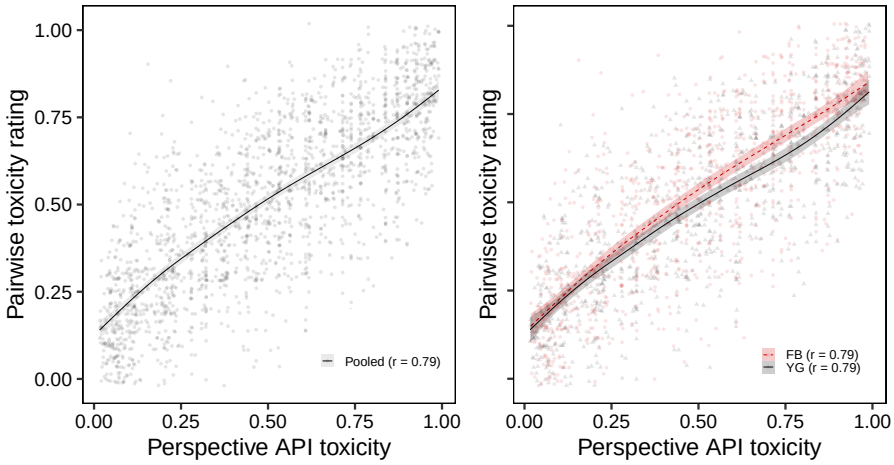
[1] The 0.3 percent of the comments with 200 words or more were excluded due to the difficulty of rating long comments in the Mechanical Turk pairwise comparison task.

[2] We audited worker performance using the methods developed by Carlson & Montgomery (2017) and identified four workers whose choices did not reflect how the comments were coded by other workers. We banned these workers from future tasks and reposted their tasks to be to be completed by other workers.

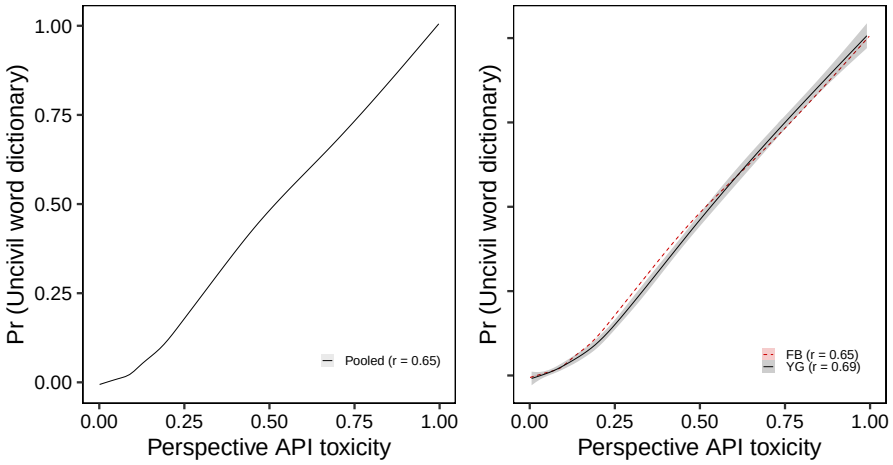
[3] The list of the features is provided in Online Appendix B. In this list, we did not include certain features that *could potentially* be used in uncivil ways but not clearly so. For example, the word "prison" can be used to say someone should be put in prison or to discuss incarceration. We excluded such features.

Figure E1: Validating the Perspective API toxicity score

(a) Perspective API vs. pairwise rating



(b) Perspective API vs. dictionary



Panel (a) includes scatterplots between the Perspective API toxicity score and crowdsourced pairwise toxicity and local polynomial fits estimating the relationship between the two measures ($N = 2,000$). The y-axis indicates the probability that a comment is chosen to be more uncivil in pairwise comparisons with 20 other randomly selected comments. Panel (b) plots local polynomial fits estimating the relationship between the API toxicity score and whether a comment includes one of the uncivil features ($N = 6,485,910$). Panel B does not plot points because the presence of an uncivil feature (plotted on the y-axis) is binary.

Crowdsourced pairwise comparison rating instrument

Instruction (Codebook)

Completing this training module qualifies you to complete Compare Comments HITs.

This task involves reading the text of two comments on news articles. Warning: The comments that you will be asked to read may include profanity or offensive language. Researchers will use your responses to better understand the "toxicity and incivility" of each comment. Your job is to read the text of both comments and select the comment that is more uncivil (or less civil).

Your performance will be monitored as you complete these HITs. We will reject all work done by workers who provide poor quality answers. Do not allow your own political opinions to influence your decisions. Your goal is to select the comment that other workers would recognize as the more uncivil (or less civil). Here are a few rules of thumb to guide you:

- Uncivil comments refer to those expressing disrespect for someone; using insulting language, profanity, or name-calling; engaging in personal attacks; and/or employing racist, sexist, and xenophobic terms
- If both comments are uncivil, pick whichever of the two comments is more uncivil
- If both comments are civil, pick whichever of the two comments is less civil
- Some comments may be meaningless. In such cases, choose whichever seems more uncivil
- Comments that use more severe name-calling or profanity are generally more uncivil
  - "He is a f***ing idiot" is more uncivil than "He is unintelligent"
- Comments that use profanity to express disrepect are more uncivil than those that use profanity to express other excitement
  - "He is f***ing stupid" is more uncivil than "He is f***ing awesome"
- Comments with more serious accusations are generally more uncivil
  - "He is a racist" is more uncivil than "He treats people unfairly"

Please read the two comments below. Your job is to read both and select the comment that is more uncivil (or less civil). We will highlight the important text and provide you with the correct answer. These will not be scored and will not count for or against your Compare Comments qualification

- "If Democrats regain control of congress. The leadership has pledged to start never-ending investigations into the current administration"
- "Pathetic losers. Democrats are so messed up. Hard to believe half the country side with these crazy people."
  - The correct answer is Comment 2. While both comments criticize Democrats, Comment 2 contains more severe name-calling.

- "Just more hate from the racist Nazi like CNN!"
- "More proof of CNN lies"
  - The correct answer is Comment 1. While both comments criticize CNN, Comment 1 is making a more serious accusation.

- "Hei"
- "She's delusional!"
  - The correct answer is Comment 2. Comment 1 does not convey any meaning whereas Comment 2 contains name-calling.

Test
You must answer those questions using only the instructions and examples provided above. You must answer at least 6 of the test HITs correctly to receive the Compare Comments qualification. You will then be eligible to the Compare Comments HITs.

Question 1
- "If it wasn't for Lies, Fraud, Deception, Intimidation and Cheating - Liberals would NEVER win at anything."
- "Im so tired of these liberals and their bullshit.They are evil and disgusting.Their talking points no longer work.They have brought a war against our Nation. Fight Back and Vote Red!" (Correct answer)

Question 2
- "Donald Trump has never read a history book in his life."
- "Your apology to that rapist will never reach my ears you orange buffoonic moron! F you!!!!!" (Correct answer)

Question 3
- "You are very misinformed"
- "You are retarded." (Correct answer)

Question 4
- "Because luv they are manipulators but too stupid to improve their own countries.... We aren't living in the dark ages you ILLEGAls how is by now you can't develop your stupid countries" (Correct answer)
- "It's mind boggling to me, especially with 12 million illegals living within your borders"

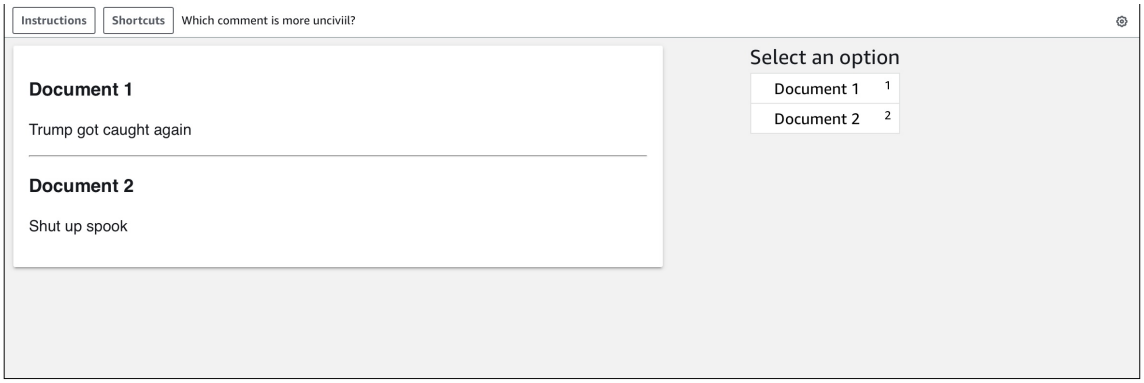
Question 5
- "Look at what these stupid stupid brain washed trump supporters say sorry we do not support hitler" (Correct answer)
- "If we let the Republicans win... then we deserve what's coming..." (Correct answer)

Question 6
- "you have to love the Republicans. Racists, bigots and homophobes. I can't believe what that party has turned into." (Correct answer)
- "If Republican leadership is going to take credit for the peaks they must take responsibility for the plunges."

Question 7
- "good"
- "PUKE" (Correct Answer)

Please continue to fully read each comment and provide your best guess of the correct answer. Your performance will be monitored as you complete more HITs. If you provide poor quality answers, all of your work will be rejected and your Compare Comments qualification will be revoked.

Classification task
Workers who passed the test were granted the qualification to complete the comparisons. The following figure provides an example screenshot of the task.



Uncivil word dictionary

“foolish”, “joker”, “lie”, “hate”, “idiot”, “stupid”, “fake”, “racist”, “hell”, “liar”, “evil”, “disgust”, “shame”, “crazi”, “moron”, “corrupt”, “fool”, “troll”, “dumb”, “loser”, “ass”, “lock”, “crap”, “socialist”, “bs”, “pathet”, “fuck”, “ridicul”, “lmao”, “cheat”, “suck”, “leftist”, “shit”, “disgrac”, “nut”, “nazi”, “snowflak”, “rapist”, “hypocrit”, “trash”, “fraud”, “traitor”, “communist”, “deplor”, “crook”, “pig”, “screw”, “smh”, “ugli”, “wtf”, “mess”, “delusion”, “nasti”, “asshol”, “witch”, “antifa”, “damn”, “freak”, “piss”, “predat”, “nonsens”, “devil”, “dirti”, “coward”, “hater”, “fascist”, “stfu”, “scum”, “jerk”, “treason”, “sheep”, “pussi”, “vile”, “brainwash”, “libtard”, “butt”, “asham”, “hitler”, “fat”, “dick”, “whine”, “hypocrisi”, “despic”, “puppet”, “bullshit”, “thug”, “narcissist”, “bigot”, “lunat”, “dumbass”, “scumbag”, “demon”, “satan”, “duh”, “hack”, “clueless”, “rat”, “creepi”, “useless”, “dump”, “arrog”, “kkk”, “sham”, “monster”, “worthless”, “sicken”, “ugh”, “tantrum”, “pervert”, “gross”, “retard”, “lmfao”, “derang”, “scam”, “dumber”, “creep”, “dummi”, “hoax”, “demonrat”, “negro”, “turd”, “pedophil”, “supremacist”, “sexist”, “misogynist”, “bastard”, “ho”, “commi”, “lazi”, “douch”, “nightmar”, “snake”, “greedi”, “cruel”, “buffoon”, “puke”, “demorat”, “vicious”, “incompet”, “whore”, “imbecil”, “anti-american”, “dumbest”, “lame”, “toxic”, “filthi”, “sucker”, “democrap”, “horrif”, “bigotri”, “jackass”, “horsefac”, “trumper”, “wth”, “dishonest”, “hooker”, “stink”, “cheater”, “brat”, “demoncrat”, “wick”, “douchebag”, “rotten”, “annoy”, “drumpf”, “spineless”, “immatur”, “fu”, “sociopath”, “chump”, “wacko”, “lier”, “cunt”, “b.s”, “marxist”, “looser”, “farc”, “trumptard”, “ffs”, “bloodi”, “sicko”, “dement”, “fraudul”, “obumm”, “brainless”, “heartless”, “dipshit”, “smdh”, “klan”, “nationalist”, “tyrant”, “asshat”, “grabber”, “tds”, “rabid”, “prostitut”, “theft”, “dementia”, “whini”, “twat”, “lowlif”, “skank”, “idioci”, “coon”, “filth”, “tramp”, “misogyni”, “af”, “redneck”, “dotard”, “bogus”, “hoe”, “psychopath”, “butthurt”, “homophob”, “unamerican”, “bum”, “coup”, “boof”, “anarchi”, “hollyweird”, “irrespons”, “classless”, “obnoxi”, “sheepl”, “friggin”, “cesspool”, “yuck”, “parasit”, “smug”, “trumpanze”, “shameless”, “dimwit”, “mf”, “asswip”, “weirdo”, “crappi”, “asinin”, “stupidest”, “tit”, “#fakenew”, “libturd”, “gtfoh”, “crock”, “maniac”, “trashi”, “gtfo”, “pimp”, “repugn”, “rubbish”, “fuckin”, “fucker”, “a-hol”, “totalitarian”, “cultist”, “adulter”, “soulless”, “whiner”, “obstructionist”, “mobster”, “maggot”, “delud”, “falsehood”, “plagu”, “horrend”, “gangster”, “nutcas”, “dupe”, “dickhead”, “dumbocrat”, “wimp”, “sissi”, “perv”, “egotist”, “slut”, “loath”, “despot”, “cock”, “goddamn”, “nitwit”, “azz”, “goon”, “fkn”, “killeri”, “clown”, “bitch”, “insan”, “communism”, “killari”, “crybabi”, “tyranni”, “anti-christ”, “psychot”, “ape”, “hillbilli”, “nutjob”