# The Causal Effects of Political Incivility in Social Media Discussions

Maxwell B. Allamong[1] Christopher A. Bail[2] D. Sunshine Hillygus[3]
*University of Houston* *Duke University* *Duke University*

Alexander Volfovsky[4]
*Duke University*

Running header: The Causal Effects of Political Incivility
Keywords: political incivility, political discussions, social media, experiments, artificial intelligence Forthcoming at the *American Journal of Political Science*

[1]Assistant Professor of Political Science, University of Houston, 424 Philip Guthrie Hoffman Hall, Houston, TX 77204, mballamong@uh.edu

[2]Professor of Sociology, Public Policy, Political Science, and Computer Science, Duke University, 347 Reuben-Cooke Building, 417 Chapel Drive, Durham, NC 27710, christopher.bail@duke.edu

[3]Professor of Political Science and Public Policy, Duke University, 140 Science Drive, 203 Gross Hall, Durham, NC 27708, hillygus@duke.edu

[4]Professor of Statistical Science and Computer Science, Duke University, 214 Old Chemistry, 415 Chapel Drive, Durham, NC 27705, alexander.volfovsky@duke.edu

This research was funded in part by the John Templeton Foundation (Award No. 62656), NSF CAREER (Award No. DMS-2046880), The Carnegie Foundation, a Facebook Foundational Research Award, and Duke University. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

# The Causal Effects of Political Incivility in Social Media Discussions

**Abstract**

Political discussions serve many important functions in democracy but increasingly occur on social media where incivility pervades. How does incivility in online political discussions impact conversational dynamics and outcomes? Using a novel research platform, we conducted a preregistered experiment testing the causal effects of incivility in online political discussions. Participants use a mobile application to access a social media feed manipulated to vary in political incivility, with discussions driven by synthetic users powered by GPT-4. Participants in the uncivil feed report less comfort sharing their views but actually created more posts. They were less likely to comment on political posts, and the content they contributed contained more uncivil features, such as profanity and insults. Furthermore, exposure to the uncivil condition led to more negative views of the out-party, while in-party ratings remained unchanged. These findings highlight how uncivil political discussions on social media simultaneously discourage open expression while fueling hostility.

**Verification Materials**: The data and materials required to verify the computational reproducibility of the results, procedures, and analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network at: https://doi.org/10.7910/DVN/L2ONWL.

The Cornell Center for Social Sciences verified that the data and replication code submitted to the AJPS Dataverse replicates the numerical results reported in the main text of this article.

Word Count: 9,819

Political discussions have long been considered a cornerstone of democracy: they offer citizens opportunities to express themselves, to encounter new perspectives, and to engage in reasoned debate (Eveland 2004; Mutz 2006; Carlson 2019). Recent work concludes that discussions held across the partisan divide can even reduce animosity and polarization (Levendusky and Stecula 2021; Combs, Tierney, Guay, Merhout, Bail, Hillygus, and Volfovsky 2023; Rossiter and Carlson 2024). Yet, much of today's political discourse occurs in online spaces such as social media where discussions are often marked by incivility (Papacharissi 2004; Coe, Kenski, and Rains 2014; Frimer, Aujla, Feinberg, Skitka, Aquino, Eichstaedt, and Willer 2023). If the context and character of online discussions matter for their outcomes, then questions arise about whether the idealized benefits of political discussion hold in online environments such as social media. We therefore ask: how does incivility shape the attitudinal and behavioral outcomes of political discussions on social media? Specifically, we evaluate the causal impacts of political incivility in a social media newsfeed on the nature and extent of engagement in discussions in that newsfeed and political attitudes more broadly.

Despite a rich literature describing the extent of incivility in online settings (e.g. Sun, Wojcieszak, and Davidson 2021; Gao, Qin, Murali, Eckart, Zhou, Beel, Wang, and Yang 2024), measuring the causal effects of social media incivility is inherently difficult given self-selection both into a particular platform and into specific conversations (Sydnor 2019; Bail 2022; Lohmann and Zagheni 2023). For example, people who want to avoid incivility may be less likely to enter (or more likely to exit) platforms without moderation, less likely engage with aggressive users, and more likely to curate their online spaces. As a result, any observed discussion dynamics and outcomes might reflect the type of people drawn to such discussions rather than the impact of incivility itself. Current experimental approaches also face limitations in estimating the causal impact of incivility. Most social media companies are unwilling to allow researchers to randomize users' exposure to uncivil content, for good reason. Researchers can create social media accounts that share uncivil content and recruit

people to follow them, but it is still difficult to know if users make a deliberate choice *not* to respond.[1] There is a rich literature that uses survey experiments to display snapshots of uncivil social media content (e.g. Hwang, Kim, and Huh 2014; Gervais 2015; Kim, Guess, Nyhan, and Reifler 2021), but such static designs typically do not capture how individuals dynamically interact with content or other users in a multi-user environment. These obstacles underscore the need for an alternative strategy to rigorously assess the effects of incivility in online political discussions.

We take a new approach to estimating the causal impacts of political incivility by conducting a preregistered experiment on a novel research platform we developed, specifically designed for controlled manipulations of social media environments (Bail, Hillygus, Volfovsky, Allamong, Alqabandi, Jordan, Tierney, Tucker, Trexler, and van Loon 2023; Allamong, Trexler, Tucker, Alqabandi, Bail, Hillygus, and Volfovsky 2024; van Loon, Katta, Bail, Hillygus, and Volfovsky 2025). The platform features a mobile application, downloaded from the Apple App or Google Play stores, that has a social media newsfeed similar to platforms like X (formerly Twitter), Bluesky, and Reddit, where uses interact under anonymous usernames to view, post, and comment on text-based content. Unbeknownst to participants, they were randomly assigned to feeds where posts and comments contained substantively similar content but were manipulated to be mostly civil or uncivil. To make the experience more dynamic and interactive, AI-powered "synthetic users" engage in the newsfeed by commenting on posts. Using embedded pre- and post-treatment surveys alongside behavioral data collected from participant interactions, we assess how exposure to these different newsfeeds shapes online engagement and broader political attitudes.

Our results show that exposure to uncivil political discussions on social media shapes people's attitudes toward self-expression, their patterns of online behavior, and cultivates out-party animus. We find that assignment to the uncivil newsfeed political reduces participants' reported comfort sharing their political views, increases animus towards out-partisans,

[1] Experimental interventions aimed at reducing incivility on real platforms acknowledge a range of other logistical and ethical challenges (e.g., Munger 2021; Ahn, Kim, Dimmery, and Munger, n.d.).

and increases the toxicity of the content they contributed to discussions on the platform. Surprisingly, those in the uncivil newsfeed were *more* likely to produce original posts, but less likely to comment on posts about politics, even for those on the same political side. Finally, we find that participants' broader perceptions of polarization, feelings of political trust, and satisfaction with democracy remain stable following this brief exposure to the uncivil newsfeed.

The argument, design, and findings presented here make three contributions. First, while political discussion is often seen as a vital component of democracy, offering opportunities for expression, engagement, and perspective-taking (e.g. Eveland 2004; Mutz 2006), we show that incivility shapes both the nature and outcomes of these interactions. In particular, incivility influences not only how people engage in political conversations but also their broader attitudes toward out-partisans. These findings are consistent with recent research showing that *how* political discussions unfold is just as important as *whether* they occur (e.g., Wojcieszak and Warner 2020; Combs et al. 2023), and suggest that the hostile tone of many online discussions could undermine the democratic benefits of such engagement. Second, in contrast to previous designs asking participants to respond to static or hypothetical content varying in civility, our novel research platform and its GPT-4-powered synthetic users offer opportunities for participants to directly engage in dynamic political discussions, if they so choose. By combining their behavior on the platform with embedded pre- and post-treatment surveys, our design allows us to examine whether political incivility on social media shapes the decision to engage, the nature of such engagement, and difficult-to-observe attitudes toward politics and the online environment. Finally, our synthetic users, programmed with civil or uncivil personas, demonstrate the broader utility of LLMs in creating dynamic online discussion environments and, more specifically, their ability to shape features of these environments that are usually hard to manipulate. Experimental designs seeking to vary speaker-level features such as incivility often require direct intervention on the part of researchers or confederates (e.g., Munger 2021; Ahn et al., n.d.) and may become costly at

scale, but we show that LLMs can offer a low-cost way to simulate groups of discussants guided through researcher prompts, offering novel insight on how discussion dynamics shape political attitudes and online behaviors.

## Background and Hypotheses

As Americans' disdain for the out-party has grown to concerning levels in recent years (Iyengar and Krupenkin 2018), many have considered how discussions across party lines could help ease political tensions (Levendusky and Stecula 2021; Amsalem, Merkley, and Loewen 2022; Combs et al. 2023; Hobolt, Lawall, and Tilley 2023). While this work rightfully emphasizes the role that political discussion plays in shaping partisan attitudes, potentially overlooked is that the promised benefits might depend on the nature and context of these interactions. For instance, much of the previous research has occurred in face-to-face settings where partisans are purposefully gathered to bridge political divides through discussion and may produce more respectful interactions as participants develop solidarity with the discussion group (Fishkin, Siu, Diamond, and Bradburn 2021; Myers 2022). By contrast, political discussions on social media are often unmoderated, anonymous, and occur among strangers, and therefore might lack the same norms or incentives for respectful engagement (Settle 2024). The group dynamic also differs, as online political discussions are not one-to-one interactions but instead occur in front of an audience that can reward, amplify, or sanction behavior. These key features suggest online discussions could lower the social costs of negative self-expression (Suler 2004; Lapidot-Lefler and Barak 2012), producing a level of incivility that constrains the positive outcomes observed in other research settings.

We thus evaluate the extent to which incivility shapes the nature and outcomes of political discussions on social media. Incivility is broadly defined as the violation of norms of decency

and respect in interpersonal communications (Papacharissi 2004; Coe et al. 2014).[2] Scholars have examined the correlates of incivility across diverse settings including higher education (Caza and Cortina 2007), healthcare (Patel and Chrisman 2020), and the workplace (Leiter, Peck, and Gumuchian 2015), often finding an association with detrimental outcomes such as psychological distress and losses of productivity. In political contexts, scholars have defined incivility as impoliteness or disrespect in the expression of one's political views, or in the criticism of the views of others, and is recognized as distinct from disagreement on policy grounds (Stryker, Conway, and Danielson 2016; Sydnor 2019).[3] Some have suggested that contentious political discussions can polarize partisan attitudes (Mutz 2007; Wojcieszak 2011; Levendusky, Druckman, and McLain 2016) and spark further uncivil engagement (Chen 2017; Kim et al. 2021). Incivility may also sour perceptions of the online environment and its participants (Gervais 2015), potentially leading to a "spiral of silence" (Noelle-Neumann 1974) where people intentionally withhold their opinions or disengage altogether (Oz 2023; Lu, Liang, and Masullo 2023).

Given this previous literature, we offer a series of hypotheses about the impact of encountering political incivility on social media. First, we expect that uncivil political discussions in a social media newsfeed will make people less comfortable sharing their political opinions (H1) and reduce their contributions to the platforms, resulting in fewer posts or comments (H2). People often consciously consider their online audience and the potential costs and benefits of sharing opinions in a discussion (Carlson and Settle 2022, 2023). Certain online settings can entail unreasonably high audience costs, such as when one hold's counter-majoritarian opinions (Allamong et al. 2024), which people may forego by simply withholding their opin-

[2]Bormann, Tranow, Vowe, and Ziegele (2022) identifies five such norms including information norms (e.g., lying or misrepresenting facts), modality norms (e.g., sarcasm, incomprehensible language), process norms (e.g., off-topic thoughts, interruptions), relation norms (e.g., name-calling, vulgarity), and context norms (in political settings, e.g., attempts to exclude opposing voices, questioning of opponent's legitimacy). Scholars have emphasized the latter two categories (e.g., Muddiman 2017) as displays of disrespect between partisans and the shutting down of cooperative discourse are hallmarks of contentious political communications (Mutz and Reeves 2005; Mutz 2007).

[3]More extreme instances of incivility may include vulgar language, hateful slurs, threats, or calls to violence. For obvious ethical reasons, we exclude these forms of incivility in our experiment.

ions (Carlson and Settle 2016; Gibson and Sutherland 2023) or by avoiding certain users or contexts (Sydnor 2019; Goyanes, Borah, and De Zúñiga 2021; Connors and Howell, n.d.).[4] Uncivil online environments may similarly raise the costs of sharing opinions, promoting discomfort and discouraging contributions to the newsfeed.

When people do engage, we expect those exposed to political incivility are themselves likely to be more uncivil in the content they produce (H3). We point to several possible reasons such a pattern may arise. First, incivility can arouse emotions such as anger, fear, or disgust (Nugier, Niedenthal, Brauer, and Chekroun 2007; Wang and Silva 2018; Gervais 2019; Kim and Kim 2019) or challenge one's social status (Masullo Chen and Lu 2017), which might influence the tone of one's response. Second, individuals may mirror the incivility they observe (Gervais 2015) or update their beliefs about norms of acceptable behavior to allow for uncivil expressions (Brady, McLoughlin, Doan, and Crockett 2021; Shmargad, Coe, Kenski, and Rains 2022). Kim et al. (2021) offer support for this expectation in a static survey experiment showing that participants displayed more toxicity in their hypothetical responses to a Facebook post when it was accompanied by toxic comments.[5] Conversely, Yeomans, Minson, Collins, Chen, and Gino (2020) show that messages signaling receptivity to cooperation (a marker of civility) reduce the likelihood of future interpersonal conflict. From this work, we expect that encountering incivility in political discussions on social media will similarly motivate incivility in users' own posts and comments, even if they share opinions less often.

Next, we consider how incivility shapes both affective evaluations of other partisans and perceptions of ideological polarization among typical Democratic or Republican voters.

[4]A defining feature of social media is its ability to connect unfamiliar individuals. Yet, interacting with strangers may fundamentally shape self-expression by incentivizing toxic behavior due to the absence of established social norms or encouraging self-censorship driven by fear of backlash.

[5]Our platform and design offer several advantages over the static survey experiments often used in communication research. First, participants are immersed in a dynamic social media environment and directly respond to incivility rather than passively viewing it. Second, the group discussion format creates the sense of an observant audience which may shape behavior. Third, static designs typically elicit immediate emotional reactions to incivility but not downstream effects such as shifts in engagement or interaction patterns over time.

Starting with affect, we expect that incivility in online political discussions will reduce favorability of both in-party (H4a) and out-party voters (H4b). Previous work has largely focused on how incivility may generate animus toward out-partisans (Rathje, Van Bavel, and Van Der Linden 2021; Goel and Merkley 2025) or inflate the gap between in- and out-party evaluations (i.e., affective polarization; Suhay, Bello-Pardo, and Maurer 2018; Skytte 2021). However, mounting evidence suggests that incivility from co-partisan media sources (Druckman, Gubitz, Lloyd, and Levendusky 2019), politicians (Frimer and Skitka 2018) and discussants (Goel and Merkley 2025) may also harm assessments of the in-party. This is because incivility is, by definition, a violation of communication norms (Papacharissi 2004), even when perpetrated by the political in-group. While some may tolerate a greater degree of incivility from the in-group before distributing punishment (Skytte 2022), evaluations of in-partisans should nevertheless be lower when exposed to uncivil political discussions on social media.

We also examine how political incivility on social media influences perceptions of polarization. If incivility amplifies pre-existing negative beliefs or homogenizes perceptions of the out-party, the uncivil condition might exaggerate perceptions of polarization. This expectation is grounded in literature suggesting that partisans generalize negative impressions to the entire out-group (Ahler 2014; Yang, Rojas, Wojcieszak, Aalberg, Coen, Curran, Hayashi, Iyengar, Jones, Mazzoleni, et al. 2016). Online incivility could also prime the belief that out-party members are ideologically extreme (Iyengar, Sood, and Lelkes 2012; Ahler and Sood 2018; Levendusky and Malhotra 2016). Exposure to incivility may reinforce these beliefs by increasing the salience of ideological extremism and strengthening the belief that partisans, but especially out-partisans, hold radical political views. Evidence from Hwang et al. (2014) supports the idea, showing that online incivility can reinforce such stereotypes and prompt people to see partisans as more ideologically distant. Thus, we hypothesize that political incivility heightens perceptions of polarization (H5).

Lastly, incivility in political discussions on social media may negatively impact broader

attitudes toward the political system such as trust in government (H6) or satisfaction with democracy (H7). Previous research has focused on social media's implications for political polarization, but many have questioned its larger impacts on democratic attitudes (e.g., Ceron and Memoli 2016; Ognyanova, Lazer, Robertson, and Wilson 2020; Guess, Malhotra, Pan, Barberá, Allcott, Brown, Crespo-Tenorio, Dimmery, Freelon, Gentzkow, et al. 2023; Allcott, Gentzkow, Mason, Wilkins, Barberá, Brown, Cisneros, Crespo-Tenorio, Dimmery, Freelon, et al. 2024; Bøggild and Jensen 2024). Incivility could drive these effects if exposure to hostile online exchanges heightens the perception that partisans are uninterested in constructive dialogue, that they deliberate in bad faith, or that the political process is fundamentally broken. Similar dynamics have been observed in contexts such as cable news (Forgette and Morris 2006; Mutz and Reeves 2005; Mutz 2007) or in government (Skytte 2021), where incivility has been tied a loss of political trust and perceived legitimacy of the opposition (Van't Riet and Van Stekelenburg 2022). By acting as a microcosm of the larger political process and reflecting the type of hostility that now exemplifies American politics, uncivil political discussions on social media, by this reasoning, could reduce trust and satisfaction with democracy.

# Research Design

We evaluate the effects of political incivility on social media by conducting a preregistered experiment on a research platform we developed called the *Social Media Accelerator* (SMA).[6,7] This platforms allows for the creation and customization of a social media experience to understand how newsfeed features and affordances (van Loon et al. 2025), partisan balance (Allamong et al. 2024), or, in this case, the civility of political content and discussion shape people's online behaviors and political attitudes. Participants access the newsfeed by downloading our *Spark Social* (Figure 1) mobile application from the Apple App Store or Google Play Store. The newsfeed shares many features of existing social media platforms

[6]Pre-registration form available at: https://doi.org/10.17605/OSF.IO/WGV2X.

[7]Information required for disclosure in the AAPOR Code of Professional Ethics & Practice available in SI A.8 (p.10).

such as X (formerly "Twitter"), Facebook, Threads, or BlueSky where users can read, post, and react to (like/dislike) online content, as well as converse through a comments section. However, unbeknownst to participants, the posts and comments that pre-populate the news-feed were crafted (and pretested) by our research team, and the other users interacting in the feed are synthetic users guided by researcher-generated prompts and powered by OpenAI's GPT-4 large language model.[8,9]

The key innovation of our platform is the ability to evaluate the causal impact of social media features and contexts on user attitudes and online behaviors within a dynamic group environment. This ability is important for studying political incivility which, as we argue, may shape both how people express themselves and whether they choose to engage at all. Observational analyses of social media platforms capture only those who join uncivil online spaces and the opinions they post. Our platform therefore offers three advantages to gaining insight on the causal effects of political incivility online. First, the respondent experience closely mirrors existing social media environments where users can create, view, and react to online content. This provides a more naturalistic setting that we hope feels familiar to our participants, though we acknowledge that participants were nevertheless aware they were testing a new and unfamiliar platform, could lead to more cautious behavior as participants learn the norms and functionalities of the online environment. Second, respondents were exposed to anonymous and unfamiliar others—conditions similar to X, BlueSky, and Reddit which may lower the social costs of hostile expressions (Suler 2004; Lapidot-Lefler and Barak 2012). These conditions also represent an important contrast to previous studies on in-person political discussions (e.g. Fishkin et al. 2021; Levendusky and Stecula 2021) or among peers (e.g., Carlson 2022) where threats of social sanction may inhibit such hostility. Finally, our

[8]Study materials were pre-tested to ensure we successfully manipulate the level of incivility across our two conditions. See SI A.6 (pp.6-9) for more details.

[9]While all users in a participants' environment were powered by artificial intelligence, a sizable majority (65%) of participants reported encountering either no bots or only one or two bots in the feed (see Figure S5, SI B, p.17). Similar patterns have been found in previous studies on the SMA platform (e.g., Allamong et al. 2024; van Loon et al. 2025), giving us confidence that our synthetic users were largely successful in simulating human behavior.

twelve-minute exposure window provided participants time to engage and form impressions within a dynamic newsfeed, while still balancing realism and experimental control. Although shorter than cumulative real-world exposure, it approximates the length of typical social media sessions (Kooti, Subbian, Mason, Adamic, and Lerman 2017) such as when waiting at the doctor's office or riding the bus, while also attempting to minimize respondent fatigue.[10]

In this study, we recruited a sample of 1,462 U.S. adults from the Prolific platform to "share their thoughts about a new social media platform."[11] Summary statistics for the sample, including a comparison to the U.S. general population and to frequent social media users, is available in Tables S2-S4 (SI B, pp.13-15).[12] We randomly assign each participant to a newsfeed in the app where we manipulate the civility of posts as well as the personas of our synthetic users who make comments, as we describe below. Our design does not include a control group with no exposure to social media discussions, as our focus is on the comparison of civil and uncivil discussions rather than social media exposure more broadly. Thus, while we can estimate the causal effect of tone in online political discussions, we cannot assess whether exposure to political discussions itself influences behavior or attitudes. Using platform behavioral data and embedded pre- and post-treatment surveys, we examine how political incivility in the newsfeed shapes peoples' online behaviors and political attitudes.

## Study Procedure & Newsfeed Experience

The study proceeded as follows (shown in Figure S1, SI A, p.3): participants from Prolific follow the link in our recruitment ad to a screening survey delivered online via Qualtrics; the recruitment ad asked people to "help us test a new social media platform" by downloading the

[10]Kooti, Subbian, Mason, Adamic, and Lerman's (2017) analysis of Facebook data reveals that 73% of user sessions on the platform are less than 15 minutes in length. Future research should consider longer time-frames to understand how effects unfold across days, weeks, or more extended periods such as entire electoral cycles.

[11]Recruitment materials available in SI A.1 (p.2).

[12]Table S4 (SI B, p.15) uses the 2024 ANES Preliminary Release to compare our sample to the U.S. general population, to frequent social media users, and to those who report posting political content on social media. Our sample is largely similar to the U.S. in terms of race and partisanship, but also younger, more feminine, more educated, and reports using more social media platforms. Our sample also appears less politically interested and less affectively polarized (lower in-party, higher out-party thermometer ratings) compared to those who post political content on social media. These comparisons are provided for descriptive context only, as our analysis is intended to identify causal effects rather than make population-level inferences.

*Spark Social* mobile application. The screening survey gathers participant's consent, confirms their eligibility for participation, and collects demographic items and political attitudes.[13] At the end of the screening survey we provide information on downloading and accessing *Spark Social* including an invite code to be entered on the app (far left of Figure 1).[14] After downloading the app and entering the invite code, participants create a profile by entering a display name and selecting an anonymous animal avatar (middle left of Figure 1). Next, they complete a Qualtrics survey within the app where we measure pre-treatment outcomes and provide instructions for interacting with the newsfeed (middle right of Figure 1). Participants are told they "have been randomly selected to join an ongoing conversation about energy, climate, and the environment." Visual guides demonstrate how to post, comment, like, and dislike content in the newsfeed, at which point participants are told they have 12 minutes in the newsfeed to "look around and interact as they wish."[15] In other words, participants are able to passively consume the newsfeed since such "lurking" is common on social media. A post that includes a "bonus code" is included to incentivize attention to the newsfeed.

[Figure 1 about here]

Next, participants are automatically redirected to the newsfeed (far right of Figure 1), randomly manipulated to be either mostly civil or uncivil in its content and in the personality of the synthetic users engaging throughout the feed (described in next section). The newsfeed features a scrollable stream of posts attributed to our synthetic users, and to which our participants can contribute their own posts and comments. Each post features three buttons that participants can engage with including a "like" button (thumbs-up) to express approval, a "dislike" button (thumbs-down) to express disapproval, and a comment button (speech bubble) to view and write comments. Clicking the like or dislike buttons increases the count shown on the post, and clicking the comment button reveals existing comments on the post

[13] Survey items available in SI A.9 (pp.10-12).

[14] Eligibility criteria include being 18 years of age or older, residing within the United States, owning an Apple iPhone or Android device, being willing and able to download the *Spark Social* mobile application, answering questions about one's self, and passing attention checks.

[15] Full instructions available in SI A.3 (pp.2-3).

and offers a text box to type one's comment. Newsfeeds are pre-populated with 12 posts and 12 comments, attributed to the other (synthetic) users, created by the research team with 4 additional posts entering the newsfeed at scheduled intervals after a participant begins the newsfeed experience.[16] Each feed includes 5 Republican and 5 Democratic synthetic users capable of dynamically commenting on posted content. These comments are generated in response to participant activity in the feed (e.g., posting, commenting), as well as in regular intervals (65 seconds) in order to simulate activity on the platform regardless of whether participants themselves are engaging in the feed.[17] Participants observe activity of the synthetic users through a comment count presented on each post and also through in-app notifications when that activity is triggered by participants posting or commenting on the platform. Importantly, each participants' newsfeed environment is self-contained, meaning they do not see posts from other study participants and the specific comments generated by the synthetic users will be unique to each individual feed (governed by the prompt engineering protocols described in the next section). The mean (median) number of dynamic comments generated by the synthetic users in a newsfeed session was 13.2 (12). For more descriptive detail, see Figure S6 (SI B, p.17).

After a participant's 12-minute newsfeed experience has ended, they are automatically redirected to a short (∼5 minute) Qualtrics survey within the app where we measure post-treatment outcomes and provide instructions for collecting compensation.[18] Overall, attrition was low: of the 1,593 participants who completed the screening survey, 1,462 completed all stages of the study, yielding a final completion rate of 91.8%. Specifically, 92 dropped out during the download phase; 10 began but did not complete the pre-treatment survey; 24

[16]The scheduled posts enter the newsfeed at 2, 4, 8, and 10 minutes into a respondent's experience. As Table S1 (SI A.5, p.6) shows, other features of the researcher-generated posts and comments are held constant across conditions including the post order, timing, number of likes/dislikes, and user attribution of content.

[17]SI A.4 (pp.4-5) details the steps in generating synthetic user comments in response to participant activity and in fixed intervals.

[18]Participants were compensated $7.50 for their participation with a potential $1.00 bonus for correctly entering a secret code ("ACE") shown in two posts from the *Spark Social* account in the newsfeed; approximately 83% of participants were awarded this bonus payment.

attrited during the newsfeed experience; and 5 did not complete the post-treatment survey.[19] We removed a single respondent who was flagged on two or more data quality indicators as specified in our pre-registration, which included survey speeding (completing the pre-treatment survey in $< \frac{1}{3}$ the median time), incoherent or blank responses to open-ended items, or a mismatch between reported state and ZIP code. This leaves a final analytical sample of $N = 1{,}461$, with $N = 734$ assigned to the civil condition and $N = 727$ to the uncivil condition.[20]

## Manipulating Incivility

The manipulation of incivility in the newsfeed occurs in two general areas discussed here and shown in Figure 2: (1) in the researcher-generated content that pre-populates or is scheduled to enter the feed, and (2) in the personas of our AI-powered synthetic users engaging dynamically with participants and other synthetic users in the newsfeed. Starting with the researcher-generated content, we manipulate incivility in the scheduled and pre-populated posts and comments about the assigned topic of energy, climate, or the environment. Reflecting the variability of actual social media (Gao et al. 2024), six of the eight pre-populated political posts are uncivil in the uncivil condition, with four additional uncivil posts entering the feed as the session progresses (and vice-versa for the civil condition). Both newsfeeds have 12 pre-populated *comments* distributed across six posts, with 10 of these comments in the uncivil newsfeed expressing a political opinion in an uncivil manner. Half of these posts and comments express opinions typically attributed to Democrats (e.g., supports green energy, opposes fossil fuels) and the other half to Republicans (e.g., opposes green energy, supports fossil fuels). Also of note is our inclusion of four non-political posts—two about chicken wings and baseball and two promoting a "bonus code" to participants—that remain fixed across the conditions. These posts offer variety in the type of content shown in the feed and also offer a sense of realism that other users are testing the platform. Furthermore, this

[19] See SI C (pp.19-20) for additional details on study attrition.

[20] Table S6 in SI A (p.16) shows these treatment arms as balanced on key demographics, political characteristics, and pre-treatment outcomes.

mix of content offers an opportunity to examine whether exposure to incivility in the newsfeed shapes the decision to engage with political versus non-political content. A schematic of the entire newsfeed is shown in Table S1 (SI A.5, p.6). We emphasize that, out of ethical concerns, the incivility treatment is quite mild—rude and disrespectful, but far short of harassment. For instance, the researcher-generated content purposefully excludes instances of hate speech, threats, slurs, or any form of abusive language. Safety features from both OpenAI and the SMA platform, such as a banned words list and a filter for content with a Perspective API toxicity score greater than 0.95, also ensure such content is not generated dynamically by synthetic users on the platform.

[Figure 2 about here]

While there is no universally-accepted measure of incivility to guide the creation of this researcher-generated content, scholars often rely on qualitative indicators of politeness or disrespect (Mutz and Reeves 2005; Mutz 2007; Sydnor 2020) or apply quantitative tools such as Perspective API's "Toxicity" measure (Wulczyn, Thain, and Dixon 2017) to classify text (e.g., Hopp, Vargo, Dixon, and Thain 2020; Kim et al. 2021; Ahn et al., n.d.).[21] We use a combination of these approaches to assess our researcher-generated posts and comments for this study, starting with a defined set of qualitative criteria inspired by Muddiman's (2017) "personal-level" and "public-level" incivility (or Bormann, Tranow, Vowe, and Ziegele's 2022 "relation" and "context" norms) that requires all researcher-generated uncivil content to contain at least two of the following features: insult or name-calling, stereotyping, swearing, capitalization or exclamation points for emphasis (i.e., shouting), attributing harm to political opponents, shutting down of political debate, question or failing to accept the legitimacy of other's opinions or motives, or the vilification of elites or issue-position holders. From any uncivil content generated by the research team, we also created a corresponding civil version

[21] Perspective describes "toxicity" as "a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion." Some scholars (e.g., Frimer et al. 2023) explicitly label this measure as "incivility," though we retain the "toxicity" label for clarity. Later we apply additional tools from Perspective to measure features such as insults, profanity, and threats, each of which may be sufficient but not necessary for language to be considered toxic.

that appears in the opposite treatment arm, helping to ensure that the issue position of a post or comment is held constant across the conditions while the civility is manipulated.[22] We validate this qualitative approach by measuring the toxicity of the posts and comments with the Perspective API, confirming little to no toxicity in the civil texts ($M \approx 0.01$) and a moderate degree of toxicity ($M \approx 0.45$) in the uncivil texts. The extent of toxicity is also roughly equal across the partisanship of the posting account, ensuring that partisanship is not a confounder. Human evaluations of the posts from workers on the CloudResearch Connect platform offer further assurance that we successfully manipulated incivility in our researcher-generated content.[23]

The second manipulation of incivility occurs in the 'persona' tags used in the prompting protocols for our LLM-powered synthetic users. These prompts were carefully engineered to guide the commenting behavior of our synthetic users, relying on a unique set of 'tags' for each user that informs the model of a user's online username, bio, persona, and political opinions (including the user's partisanship and issue positions).[24] Both newsfeeds contain five synthetic users tagged as Democrats and five as Republicans, but in the uncivil (civil) newsfeed eight of these users (80%; four Democratic, four Republican) are assigned persona tags of the types of people likely to engage in uncivil (civil) political discussion while the two remaining (20%, one Democratic and one Republican) are tagged as the opposite type. For example, uncivil synthetic users were tagged as "toxic," "disrespectful," or "impolite" while civil synthetic users were tagged with more mundane qualities like "laid-back," "respectful," and "reserved." Participants did not see these words or any indication of the tags, they only viewed the resulting comments produced by the synthetic users. We arrived at these

[22]While we attempted to manipulate only these uncivil features in our researcher-generated content, we acknowledge the possibility that other linguistic features (e.g., expressed certainty) may sometimes co-vary with our manipulation. This reflects our effort to present incivility in a realistic and contextually appropriate manner, though we recognize it also limits our ability to isolate the precise linguistic elements responsible for the observed effects.

[23]See SI A.6 (pp.6-9) for more on pre-testing of study materials.

[24]Our prompts follow a Role-Task-Format (RTF) framework where the model is informed of the personal qualities it is to assume (role), the action it is to perform (task), and details for the structure of the model's output (format). See SI A.4 (pp.4-5) for more on prompting.

sets of tags following a careful prompt engineering process that included extensive internal qualitative testing, as well as quantitative validation through both the Perspective API toxicity measure and human evaluations via CloudResearch Connect. SI A.6 (pp.6-8) shows the degree of incivility to be slightly less, but nevertheless detectable, in this LLM-generated content compared to researcher generated content, and at roughly equal rates across our Democratic and Republican users.

With these manipulations in mind, treatment in our design is the totality of the experience of being assigned to either the *civil* or the *uncivil* newsfeed. Defining treatment in this way is important because the exact experience of participants assigned to the same condition may vary slightly for several reasons. First, synthetic users routinely comment on the platform about once a minute, but the commenting bot is selected through a probabilistic process described in SI A.4 (pp.4-5) so users may not see the same comments appearing in their feed. Second, there is natural variation in the output of repeated calls to GPT-4 even given the exact same prompt; again our pre-tests give us confidence that, on average, our civil and uncivil users acted accordingly. And finally, events such as user comments and posts also triggered a probabilistic comment generation process described in SI A.4 (pp.4-5) so users could engage in dynamic conversations in response to their posted content. We view this type of treatment as analogous to treatments delivered in fields such as medicine and education where nuanced interactions between treatment (e.g., assigned medicine or educator) and subject (e.g., research participant or student) are also likely to occur. Our interest, as in those fields, is the totality of the experience with the research intervention, recognizing a degree of variation in the treatment. Importantly, the prompts and all content created for each respondent's treatment are captured by our platform and made available in the replication materials. A manipulation check measured after our primary outcomes gives

us confidence that our treatment was delivered.[25]

## Hypotheses & Empirical Strategy

In the empirical analyses to follow, we examine the causal effects of being randomly assigned to the uncivil feed relative to a civil feed. We estimate these casual effects by regressing each respective outcome on an indicator for assignment to the uncivil feed and interpret the substantive and statistical significance of the coefficient estimate for this indicator. Our outcome with respect to H1 is a 5-point post-treatment measure asking how comfortable people felt sharing their views about political issues on the app.[26,27] For Hypothesis 2, we operationalize content sharing to the platform with a count of the number of posts (H2a), number of comments (H2b), and number of comments on specifically out-party posts (H2c) made by the user, as well as their respective level of toxicity (H3a-c) from the Perspective API ([0,1]).[28] Attitudes toward other partisans in H4a-b are measured with standard 101-point feeling thermometer ratings of in- (H4a) and out-party (H4b) voters. Perceptions of polarization in H5 are measured by asking participants to place voters from either party on 7-point scales representing voters' beliefs about the governments' role in addressing climate change, and calculating the absolute distance between the placements. Political trust in H6 is measured with a revised item from the American National Election Studies (ANES) asking

[25]Our manipulation check assessed whether participants perceived the conversations on the platform as mostly civil or mostly uncivil. Nearly all participants (98%) in the civil condition rated the conversations as mostly civil, with only 2% perceiving them as mostly uncivil. In contrast, in the uncivil condition, approximately 40% of participants rated the conversations as mostly uncivil, while the remaining 60% rated them as mostly civil. Although the absolute level of perceived incivility was modest, the difference between conditions is large and statistically significant, indicating that exposure to uncivil content reliably shifted perceptions in the intended direction. Because our manipulation was designed to approximate the subtle and uneven character of incivility in real online discussions, where even hostile environments are not uniformly so, it is unsurprising that many participants in the uncivil condition still perceived the overall discussion as mostly civil. These patterns provide strong evidence that our manipulation effectively varied the level of incivility between the two conditions, even though the overall degree of incivility was relatively mild, as anticipated from our pre-testing.

[26]Question wording for all outcomes available in SI A.9 (pp.10-12).

[27]While our use of single-item measures may raise concerns about reliability, the specific measures we employ—such as feeling thermometers, left-right party placement items, and standard ANES items on political trust and democratic satisfaction—are widely used and validated in political science research. We also opted for single-item formats to reduce respondent burden given the overall length and complexity of the study.

[28]Hypotheses 2c and 3c flow directly from our other hypotheses but were not explicitly listed in our pre-registration plan.

how often people "can trust the government in Washington to do what is right." Finally, satisfaction with democracy in H7 is also a revised item from the ANES gauging assessments of "the way democracy is working in the United States." Note that in testing H4-H6 we use the respective (standardized) post-treatment measure as the outcome while controlling for the pre-treatment measure of the same item.[29] We report exact p-values, using one-tailed tests at $\alpha = 0.05$ for preregistered directional hypotheses and two-tailed tests at the same threshold for exploratory analyses. Our expectations are summarized in Table 1.

[Table 1 about here]

# Results

Figure 3 presents the average treatment effects (ATEs) of assignment to the uncivil newsfeed relative to the civil newsfeed on the outcomes associated with Hypotheses 1-4. Starting at the top, we find those assigned to the uncivil newsfeed reported being significantly less comfortable sharing their political views on the app compared to those assigned to the civil feed ($\hat{\beta} = -0.293, p < 0.001$). With a standard deviation of 1.3 on this outcome, this standardized effect translates to a roughly 0.4 point difference on a five-point scale between the two conditions. These findings support Hypothesis 1 that incivility makes people less comfortable sharing their views, aligning with recent research suggesting that the contentious political climate in the U.S. has made people more conscious of the potential social costs of sharing political opinions unfavorable to others (Carlson and Settle 2022; Gibson and Sutherland 2023).

[Figure 3 about here]

[29] This analytical strategy deviates from our pre-registration plan which stated we would use within-subject change in any outcomes measured both pre- and post-treatment. The reason for this deviation is that this analytical strategy offers lower variance in our estimates without introducing additional bias (Gerber and Green 2012; Blair, Cooper, Coppock, and Humphreys 2019). We nevertheless report the pre-registered outcomes in Table S10 (SI D, p.22) which confirm that the interpretation of the statistical significance of our estimates remain the same.

Given our expectation that participants would feel less comfortable sharing their views in the uncivil condition, we had also expected they would make fewer posts and comments. Somewhat surprisingly we find that those assigned to the uncivil newsfeed overall made significantly *more* posts ($\hat{\beta} = 0.072, p = 0.027$) compared to those in the civil feed.[30] Participants in the uncivil feed also made slightly more comments overall ($\hat{\beta} = 0.188, p = 0.112$), although the difference is not statistically significant.[31] However, when we look specifically at comments made on out-party posts, there is some suggestion that incivility could discourage cross-partisan interactions ($\hat{\beta} = -0.095, p = 0.262$) though the effect is not statistically significant.

Looking next at the substance of this user-generated content, we find that assignment to the uncivil feed also significantly influenced users' manner of expression, with higher levels of toxicity observed in both the posts ($\hat{\beta} = 0.371, p < 0.001$) and comments ($\hat{\beta} = 0.356, p < 0.001$) of those exposed to uncivil feed. With standard deviations of 0.1 on either outcome, the standardized effects correspond to raw differences of roughly 0.04 points on the 0–1 toxicity scale. Similar results are found when looking specifically at comments on out-party posts, which are on average significantly more toxic among those assigned to the uncivil feed ($\hat{\beta} = 0.410, p < 0.001$), corresponding to a raw difference of 0.04 points. These findings lend support to Hypothesis 3 and suggest that users are responding to the uncivil online environment by themselves becoming more hostile, further contributing to an online environment that makes people uncomfortable to express themselves freely.

The conventional wisdom is that most online toxicity comes from a small minority of highly uncivil users (Kumar, Hancock, Thomas, and Durumeric 2023), so we look descrip-

[30] While our pre-registration did not specify adjustments for multiple hypothesis tests, our results are largely robust to such corrections. Our finding regarding the effect of the uncivil newsfeed on the number of posts is the only result where statistical significance is sensitive to adjustment, with $p = 0.053$ using a Benjamini & Hochberg correction to control the false discovery rate, and $p = 0.32$ using the Bonferroni correction to adjust the family-wise error rate, which we view as overly conservative in this context.

[31] Table S5 (SI B, p.16) offers additional summary statistics on participants' posts and comments such as comparisons of content length across conditions. We show, for instance, that comments made in the civil condition contained significantly more tokens and characters on average. Figure S7 (SI B, p.18) shows the distribution of post and comment counts across the two conditions.

tively to see who on the platform was producing this uncivil content (see supplemental analyses, Table S7, SI B, p.19). We group participants into non-posters and low ($> 25^{th}$ pctl.), medium ($25^{th} - 75^{th}$ pctl.), and high ($> 75^{th}$ pctl.) toxicity posters, as well as non-commenters and low, medium, and high toxicity commenters. Across comments and posts, we find that high toxicity users are more likely to be male, report using more social media platforms, create sizably more content overall, and report greater comfort sharing their opinions than less-toxic and non-toxic users. These patterns are consistent with previous observational research finding a small number of highly uncivil participants can disproportionately shape the tone of the discussion, overshadowing more civil voices and potentially creating distortions in people's perceptions of online social norms and public opinion more broadly (Robertson, Del Rosario, and Van Bavel 2024).

Turning to the results for party attitudes, estimates at the bottom of Figure 3 offer evidence that exposure to the uncivil newsfeed cultivates animus toward out-party voters in the form of lower thermometer ratings ($\hat{\beta} = -0.085, p < 0.001$). With a standard deviation of roughly 21 points, this standardized effect translates to a raw difference of approximately 1.8 points on the 0-100 feeling thermometer, or about one-quarter of the 6.8 point drop in out-party affect observed in the ANES between 2016-2020. Attitudes toward in-party voters, however, show no evidence of change ($\hat{\beta} = 0.020, p = 0.244$). These findings suggest that the nature of online political discussions—whether civil or uncivil—plays a pivotal role in determining whether people leave the exchange feeling more positive toward the other side—as argued by previous work—or more disdainful as we show here.

The findings so far highlight the impact of incivility in shaping the nature and outcomes of online political exchanges. We next turn to some exploratory analysis to help give some context to these findings, leveraging the rich behavioral data on the platform. We start by examining how assignment to the uncivil feed shapes specific conversational features in user-

generated content.[32] Figure 4 shows for posts (triangles) and comments (circles) the average treatment effect of assignment to the uncivil feed on features including: a 'politeness' index (reversed; Yeomans, Kantor, and Tingley 2018) sentiment from the Valence Aware Dictionary and sEntiment Reasoner or "VADER" model (reversed; Hutto and Gilbert 2014), and three additional metrics from the Perspective API related to identity attacks, insults, and profanity.[33] Here we see some of the detailed ways in which incivility is shaping users' language: for instance, not only are the comments of those in the uncivil condition more likely to contain impolite features ($\hat{\beta} = 0.125, p = 0.033$), but they are also significantly more negative in sentiment ($\hat{\beta} = 0.116, p = 0.049$) and more likely to include concerning elements such as attacks on identity characteristics ($\hat{\beta} = 0.229, p < 0.001$), insults ($\hat{\beta} = 0.356, p < 0.001$), and profanity ($\hat{\beta} = 0.269, p < 0.001$). Participants' posts show relatively similar patterns with the exception of the sentiment measure (see Table S12, SI D, p.23). Somewhat reassuringly, the overall prevalence of features such as identity attacks, insults, and profanity is quite low (see Table S16, SI D, p.25), though in relative terms they are substantially more prevalent in the content generated by users in the uncivil condition by about 0.2-0.3 standard deviations. This analysis underscores how uncivil environments influence not just the volume but also the tone of online political discussion, specifically influencing users' linguistic patterns in ways that contribute to a more hostile and less comfortable conversational space.

[Figure 4 about here]

We also explore how incivility changes commenting behavior across different types of content to see if uncivil political discussions lead people to avoid political content, especially from out-partisans (Skoric, Zhu, Koc-Michalska, Boulianne, and Bimber 2022). In the left panel of Figure 5 we plot the proportion of respondents who commented on at least one political post from an in-partisan source (left bars), an out-partisan source (middle bars),

[32] Our pre-registration states that we may perform exploratory analyses of content shared to the platform, but did not outline specific hypotheses. We report 95% intervals and p-values from two-tailed tests where appropriate for the exploratory analyses performed in Figures 4-5 and Figures S9-S10 (SI D, p.26)

[33] Our approach to creating the politeness index follows Combs et al. (2023) where all 'politeness' features for a given piece of content (text or post) are standardized and then an averaged.

or on a non-political post from an apolitical source (right bars). The figure compares users in the civil (gold) and uncivil (purple) conditions. Here we see that participants assigned to the uncivil feed where less likely to comment on posts about politics from either in-partisan ($\Delta = -0.063, p = 0.022$) or out-partisan profiles ($\Delta = -0.062, p = 0.026$, see Table S17, SI D, p.25) by about 6 percentage points, but were somewhat more likely to comment on a non-political post ($\Delta = 0.037, p = 0.114$) by about 3 percentage points.[34] These results suggest that incivility not only leads people to avoid politics but possibly to divert their attention to other forms of content.

[Figure 5 about here]

Additional exploratory analysis examines the diffusion of incivility through the newsfeed (Brady, Wills, Jost, Tucker, and Van Bavel 2017; Mamakos and Finkel 2023), to see if it spreads from uncivil political content to the more mundane non-political content. The right panel of Figure 5 shows the raw toxicity of users' comments across each type of content. Here we see that exposure to the uncivil feed promotes toxicity in users' comments made on posts from in-partisan ($\Delta = 0.030, p < 0.001$) and out-partisan profiles ($\Delta = 0.037, p < 0.001$) by about 3-4 percentage points, but does not promote toxicity in comments made on non-political posts ($\Delta = 0.003, p = 0.6$, see Table S17, SI D, p.25). Combined with our previous result about the types of content people engaged with, this result indicates that exposure to incivility on social media leads to a concentration of toxic behavior in political discussions, as people become less likely to engage with political content in the first place but are more toxic when doing so. These sort of complex reactions to online incivility are not easily uncovered in the typical static survey experiment which tends to measure isolated responses to a single piece of content. While useful for identifying immediate reactions to observed incivility, such designs do not capture how people behave within a broader social environment. By contrast, our study embeds participants directly in a discussion group, allowing us to observe not only

[34] P-values from two-tailed tests reported for our exploratory analyses.

how they respond to incivility, but also how it shapes their behavior and attention across other posts, comments, and topics in the feed.

We also explore the temporal dynamics of incivility, examining the sequencing of uncivil comments (see Figure S9, SI D, p.26). In both conditions, participants' comments were somewhat more toxic when responding to posts that were already uncivil, suggesting a tendency to match the tone of the initiating post. In our second analysis, we analyze the toxicity of comments across the 12-minute newsfeed session to examine whether toxic behavior, particularly in the uncivil condition, changes over time. Figure S10 (SI D, p.26) shows that comment toxicity in the uncivil condition began at a higher level than in the civil condition and remained elevated throughout the session, but also increased over time in both conditions. This pattern suggests that while initial exposure to an uncivil environment leads to a higher baseline of toxicity, prolonged exposure to either environment is associated with a gradual escalation in toxic expression. Together, these analyses suggest that uncivil environments not only draw participants into engaging with and reinforcing incivility when it appears, but can also encourage them to inject incivility into otherwise civil exchanges. Furthermore, the steady increase in toxicity across both conditions points to a broader dynamic in which extended participation in online political discussions, regardless of the starting tone, can gradually erode civility, amplifying more toxic forms of expression over time.

Returning to our final hypotheses, we test the popular wisdom that exposure to incivility will have spillover impacts on broader attitudes toward the political system. These results are shown in Figure 6. Starting with H5 at the top, we find no evidence that assignment to the uncivil feed had an affect on people's perceptions of ideological polarization between opposing voters ($\hat{\beta} = 0.007, p = 0.428$). Trust in government (H6) shows similar null results as assignment to the uncivil newsfeed has no distinguishable effect on this outcome ($\hat{\beta} = 0.015, p = 0.271$). Then at the bottom of Figure 6 we see that satisfaction with democracy moves in the expected direction as assignment to the uncivil newsfeed slightly decreases

democratic satisfaction ($\hat{\beta} = -0.026, p = 0.178$); however, this estimate does not reach statistical significance and thus H7 also lacks support in this study.

[Figure 6 about here]

# Discussion

Many have suggested that political discussions are crucial to sustaining democratic societies by helping to foster mutual understanding and even mitigate polarization (e.g., Mutz 2006; Levendusky and Stecula 2021). However, many discussions are occurring on social media where incivility is frequently observed, raising questions about how the civility of those discussions might shape the outcomes. We evaluate the role of incivility in political discussions in a social media newsfeed using a novel experimental design where we manipulated incivility in posts, comments, and the personas of our AI-powered synthetic users who engaged in the feed. The results of this experiment demonstrate that political incivility reduces participants' reported comfort sharing their political views, increases animus for out-party voters, and increases the toxicity of their own contributions to discussions on the platform. Exploratory analyses then reveal that political incivility in the feed discourages engagement with political content from both in- and out-partisans, which is also where users tend to concentrate their toxic language. Broadly these results suggest that incivility in political discussions on social media—even in a mild form—shapes whether, where, and how people engage online, and may deepen the already high levels of out-party animus.

Our experiment found no evidence that the uncivil newsfeed affected people's attitudes toward their own party. A possible explanation is that our political incivility treatment is rather mild compared to real world platforms, which may be insufficient to move in-party support as previous work has found (Levendusky et al. 2016; Frimer and Skitka 2018; Skytte 2022). We likewise found no evidence that exposure to online political incivility moved people's broader attitudes such as their perceptions of polarization, trust in government, or satisfaction with democracy. These null findings are consistent with recent, large-scale social

media field experiments (e.g., Guess et al. 2023) which also found system-level attitudes to be mostly unresponsive to changes in the online environment. One possibility is that, to the extent these attitudes are influenced by political incivility, their effects may emerge gradually or require sustained exposure to incivility, and our exposure may be insufficient to nudge these attitudes. Future work examining a longer or more intense exposure to incivility may revisit these outcomes.

As with any experimental study, there are design limitations to consider. First, although our Social Media Accelerator platform offers a more ecologically valid setting than traditional survey experiments—allowing for precise manipulation of social media features in a dynamic, interactive environment—it remains an artificial context. Participants were specifically recruited to test a new and unfamiliar platform, interacting with strangers in a controlled setting which may not fully capture the norms, habits, and emotional investments that shape behavior on more familiar platforms in their own social networks. Nevertheless, the experience of interacting with strangers is not unusual online, and our platforms purposefully shares many features of real-world platforms while not intending to replicate any one in particular. Second, the relatively short duration of the newsfeed experience (12 minutes) constrains our ability to assess the cumulative effects of repeated exposure to incivility over time. Such prolonged exposure, such as what one might experience throughout election season, could amplify its effects in ways not captured in our design. Next, while our study examined outcomes ranging from behaviors on the platform to self-reported attitudes like political trust or democratic satisfaction, there are other plausible outcomes overlooked, including spillovers of discomfort or resentment across online platforms, or spillovers to offline interactions or political behaviors. Finally, our design included only text-based incivility, while many platforms are video-based (e.g., YouTube, TikTok), and its plausible that the observed effects could be different in such environments. Though prior research suggests people perceive greater incivility through formats such as video even when content is held constant (Sydnor 2020), future studies should explore how factors such as familiarity with a

platform, or duration and mode of exposure, condition the effects of online political incivility.

Broadly, this work advances our understanding of how incivility in political discussions on social media influences their dynamics and outcomes. Efforts to bridge political divides through discussion remain important, particularly in this era of historic polarization. However, our findings suggest not all discussion environments foster constructive interactions. In fact, uncivil social media environments may lead to unexpected and potentially harmful consequences. These results contribute to a rich body of research on the effects of incivility in political communication which gained prominence during the rise of cable news (e.g., Brooks and Geer 2007; Mutz 2015), but has since proved more challenging to study in an era of social media. While creating an online discussion environment that mimics the dynamic nature of social media and varies systematically in its incivility is a substantial task, we overcome these challenges with our novel research platform and its AI-powered synthetic users. We hope our platform and methodological approach offer a hopeful path forward for investigating other complex dynamics on social media.

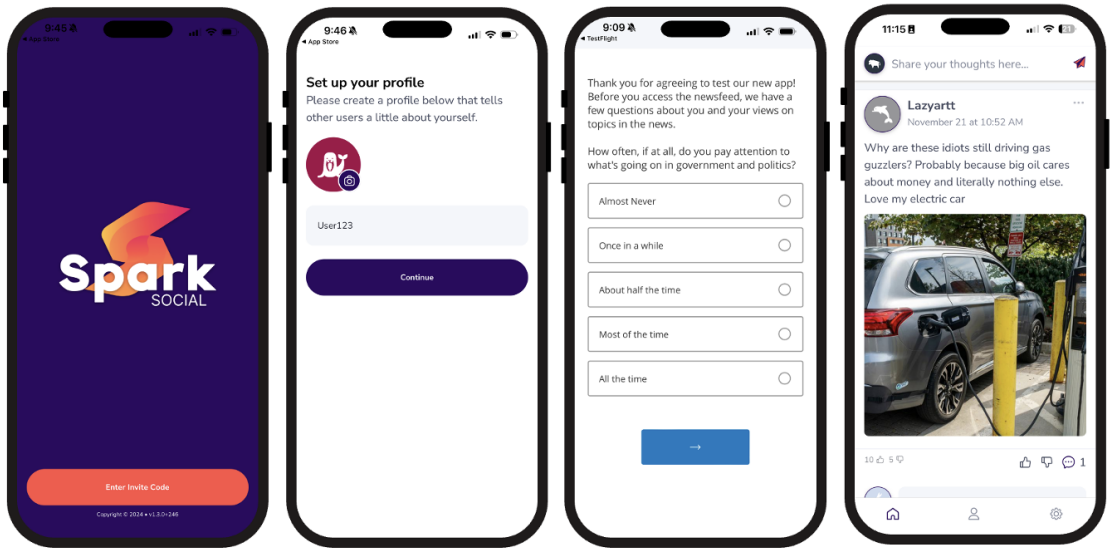



Figure 1: Screenshots from the Spark Social mobile application

Note: Figure shows Spark Social mobile application landing page (far left), profile creation page (middle left), embedded Qualtrics survey (middle right), and post in newsfeed with associated image and comments (far right).

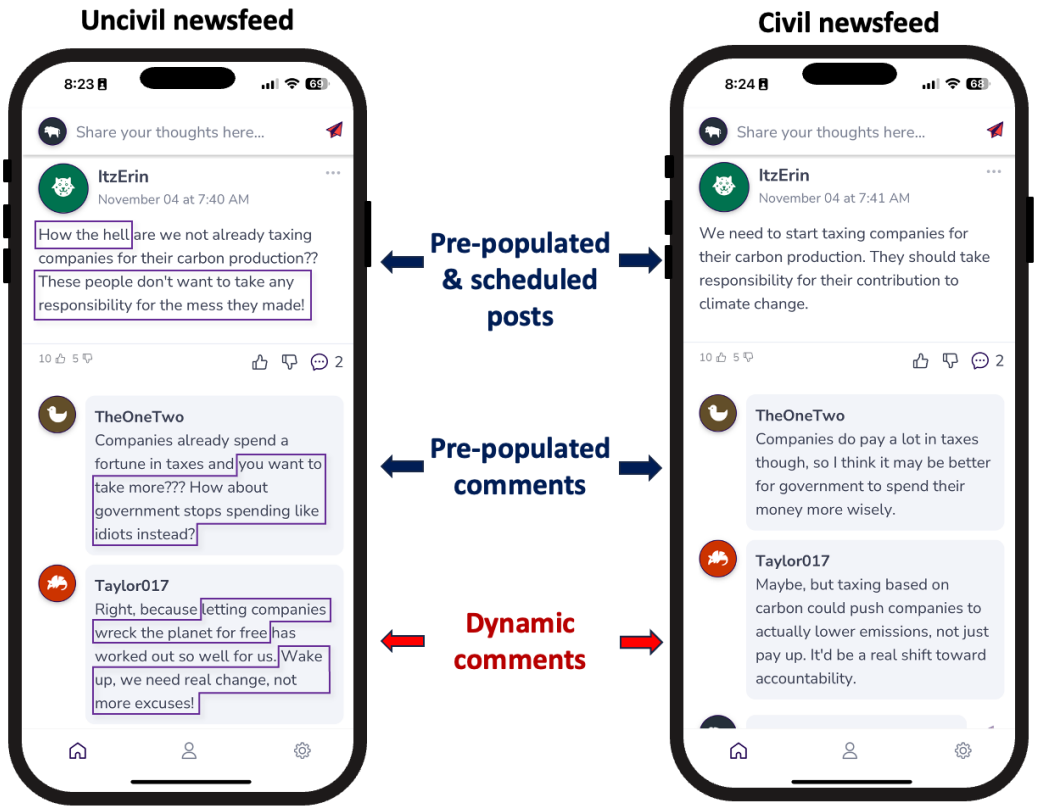



Figure 2: Examples of content from uncivil and civil newsfeeds

Note: Examples of posts and comments in the newsfeeds of the uncivil (left) and civil (right) conditions. Researcher-generated content in blue; GPT-generated comments in red. Purple boxes mark uncivil content in the uncivil condition.

Table 1: Summary of hypotheses, measures, and expected effects of uncivil newsfeed

| **Hypothesis** | **Description** | **Measure** | **Effect** |
|---|---|---|---|
| H1 | Comfort sharing political views | 5-point scale | - |
| H2a | Number of posts | Count | - |
| H2b | Number of comments | Count | - |
| H2c | Comments on out-party posts | Count | - |
| H3a | Toxicity of posts | 0–1 (Perspective API) | + |
| H3b | Toxicity of comments | 0–1 (Perspective API) | + |
| H3c | Toxicity of out-party comments | 0–1 (Perspective API) | + |
| H4a | Thermometer: in-party voters | 101-point scale | - |
| H4b | Thermometer: out-party voters | 101-point scale | - |
| H5 | Perceived polarization | 7-point interparty distance | + |
| H6 | Trust in government | 5-point scale | - |
| H7 | Satisfaction with democracy | 5-point scale | - |

Note: Signs denote expected directional effects of assignment to the uncivil (vs. civil) condition. All measures standardized in analyses except counts of posts, comments, and out-party comments in H2. Measures for H4–H7 taken pre- and post-treatment.

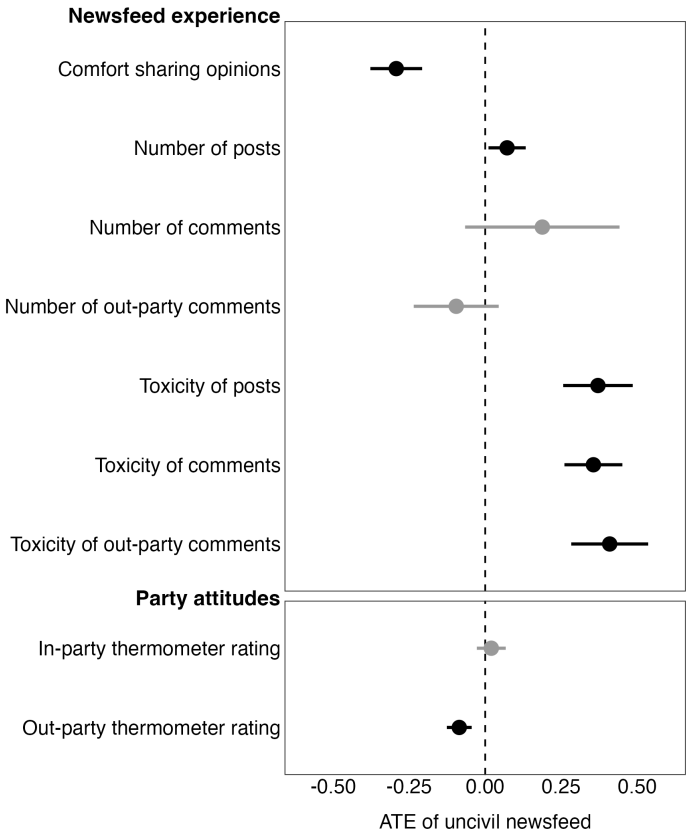



Figure 3: Average treatment effect (ATE) of the uncivil newsfeed on newsfeed experience and party attitudes

Note: 90% confidence intervals shown. Black points indicate statistical significance at $\alpha = 0.05$ (one-tailed); gray points indicate non-significant estimates. Measures for H1, H3, and H4 are standardized; H2a-c outcomes are raw counts. Estimates correspond to models in Tables S8-S9 (SI D,p.21); random-intercept models (Tables S11, SI D, p.22) yield consistent results for toxicity outcomes.

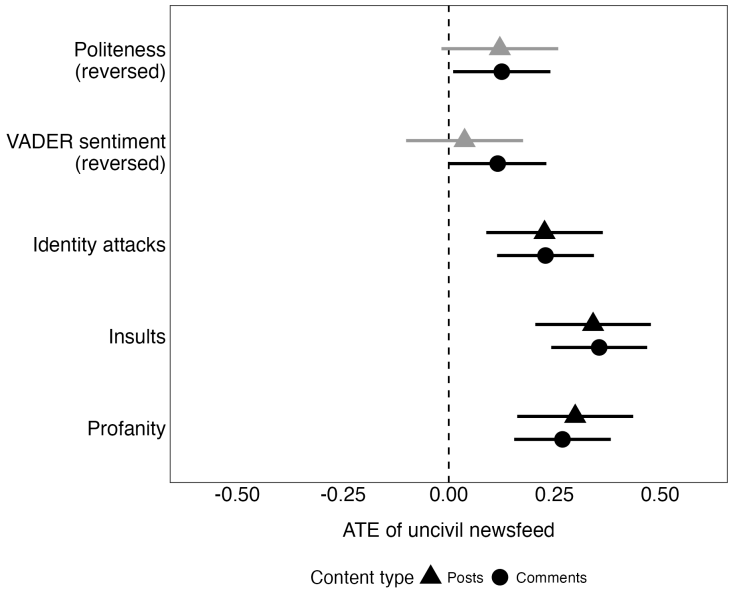



Figure 4: Average treatment effects (ATE) of the uncivil newsfeed on conversational features of users' posts and comments

Note: 95% confidence intervals shown. Black points indicate statistical significance at $\alpha = 0.05$ (two-tailed); gray points indicate non-significant estimates. Outcomes reflect participants' average use of each conversational feature across posts or comments and are standardized for comparison. Estimates correspond to models in Tables S12-S13 (SI D, p.23); random-intercept models (Tables S14–S15, SI D, p.24) yield consistent results. Raw means for all measures are reported in Table S16 (SI D, p.25).

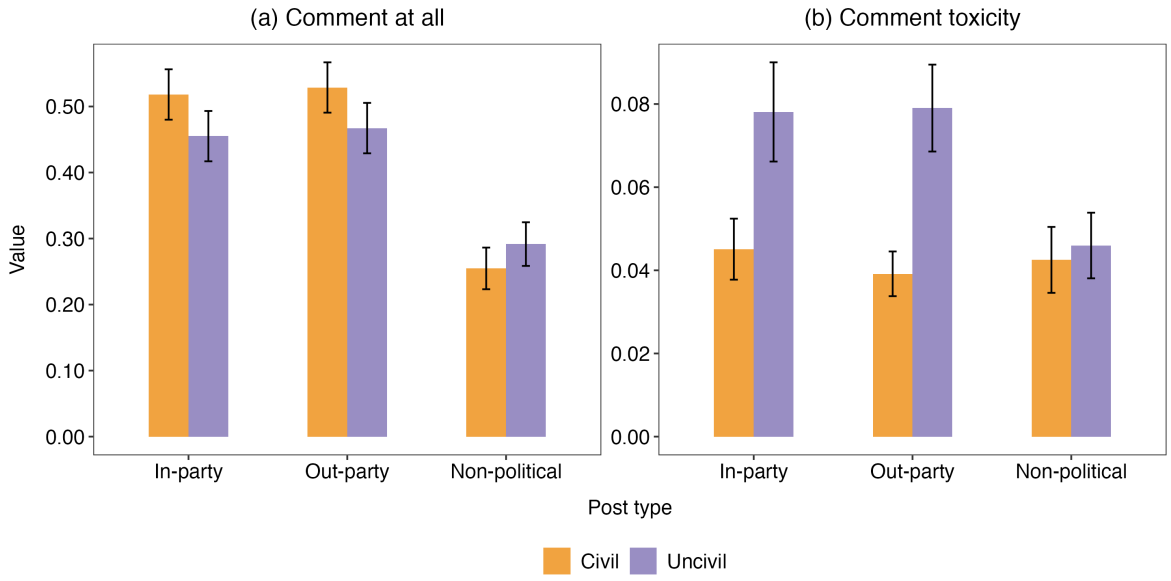



Figure 5: Commenting and comment toxicity on in-party, out-party, and non-political posts by condition

Note: 95% confidence intervals shown. Left panel shows the proportion of respondents who commented at least once by post type; right panel shows mean comment toxicity. Political posts were classified as in- or out-party based on the respondent's partisanship relative to the synthetic user who posted each message. Bonus and filler posts (non-political) were identical across conditions.

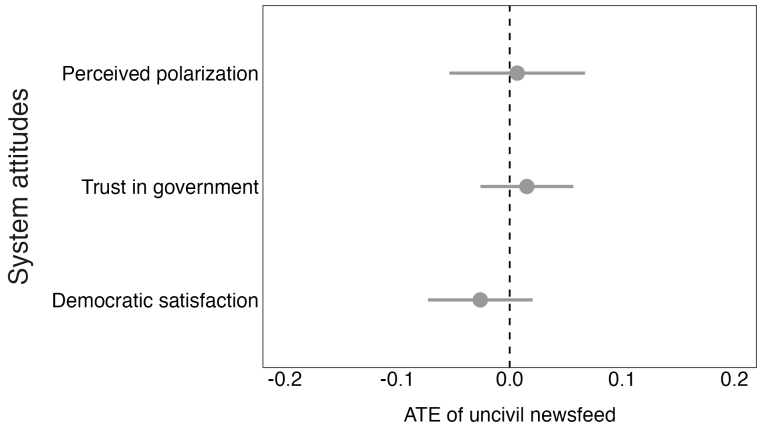



Figure 6: Average treatment effects (ATE) of the uncivil newsfeed on system attitudes

Note: 90% confidence intervals shown. Black points indicate statistical significance at $\alpha = 0.05$ (one-tailed); gray points indicate non-significant estimates. Outcomes are standardized and models control for pre-treatment values. Estimates correspond to Table S9 (SI D, p.21).

# References


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

Ahler, Douglas J, and Gaurav Sood. 2018. "The Parties in Our Heads: Misperceptions about Party Composition and Their Consequences." *The Journal of Politics* 80 (3): 964–981.

Ahn, Chloe, Sangyeon Kim, Drew Dimmery, and Kevin Munger. n.d. "Twitching, Fast and Slow." *Proceedings of the 10th International Conference on Computational Social Science (IC2S2),* https://osf.io/aqb7j/files/248mx.

Allamong, Maxwell B., Andrew Trexler, Tina Tucker, Fatima Alqabandi, Christopher A. Bail, D. Sunshine Hillygus, and Alexander Volfovsky. 2024. "Outnumbered Online: An Experiment on Partisan Imbalance in a Dynamic Social Media Environment." *OSF Preprints,* https://osf.io/preprints/osf/tygec.

Allcott, Hunt, Matthew Gentzkow, Winter Mason, Arjun Wilkins, Pablo Barberá, Taylor Brown, Juan Carlos Cisneros, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, et al. 2024. "The effects of Facebook and Instagram on the 2020 election: A deactivation experiment." *Proceedings of the National Academy of Sciences* 121 (21): e2321584121.

Amsalem, Eran, Eric Merkley, and Peter John Loewen. 2022. "Does Talking to the Other Side Reduce Inter-party Hostility? Evidence from Three Studies." *Political Communication* 39 (1): 61–78.

Bail, Christopher A. 2022. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing.* Princeton, NJ: Princeton University Press.

Bail, Christopher A., D. Sunshine Hillygus, Alexander Volfovsky, Max Allamong, Fatima Alqabandi, Diana ME Jordan, Graham Tierney, Christina Tucker, Andrew Trexler, and Austin van Loon. 2023. "Do We Need a Social Media Accelerator?" *SocArXiv,* https://doi.org/10.31235/osf.io/ucfbk.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Use Change Scores or Control for Pre-Treatment Outcomes? Depends on the True Data Generating Process." *DeclareDesign Blog Post,* https://declaredesign.org/blog/posts/use-change-scores-or-control.html.

Bøggild, Troels, and Carsten Jensen. 2024. "When politicians behave badly: Political, democratic, and social consequences of political incivility." *American Journal of Political Science* 69 (3): 1064–1081.

Bormann, Marike, Ulf Tranow, Gerhard Vowe, and Marc Ziegele. 2022. "Incivility as a Violation of Communication Norms—A Typology Based on Normative Expectations toward Political Communication." *Communication Theory* 32 (3): 332–362.

Brady, William J, Killian McLoughlin, Tuan N Doan, and Molly J Crockett. 2021. "How social learning amplifies moral outrage expression in online social networks." *Science Advances* 7 (33): eabe5641.

Brady, William J, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. "Emotion shapes the diffusion of moralized content in social networks." *Proceedings of the National Academy of Sciences* 114 (28): 7313–7318.

Brooks, Deborah Jordan, and John G Geer. 2007. "Beyond Negativity: The Effects of Incivility on the Electorate." *American Journal of Political Science* 51 (1): 1–16.

Carlson, Taylor N. 2019. "Through the Grapevine: Informational Consequences of Interpersonal Political Communication." *American Political Science Review* (Cambridge,UK) 113 (2): 325–339.

———. 2022. *Through the Grapevine: Social Transmitted Information and Distorted Democracy.* Chicago, IL: University of Chicago Press.

Carlson, Taylor N, and Jaime E Settle. 2016. "Political Chameleons: An Exploration of Conformity in Political Discussions." *Political Behavior* 38 (4): 817–859.

Carlson, Taylor N, and Jaime E Settle. 2022. *What Goes Without Saying: Navigating Political Discussion in America.* Cambridge, UK: Cambridge University Press.

———. 2023. "Freedom of Expression in Interpersonal Interactions." *PS: Political Science & Politics* 56 (2): 245–249.

Caza, Brianna Barker, and Lilia M Cortina. 2007. "From Insult to Injury: Explaining the Impact of Incivility." *Basic and Applied Social Psychology* 29 (4): 335–350.

Ceron, Andrea, and Vincenzo Memoli. 2016. "Flames and Debates: Do Social Media Affect Satisfaction with Democracy?" *Social Indicators Research* 126:225–240.

Chen, Gina Masullo. 2017. *Online Incivility and Public Debate: Nasty Talk.* New York, NY: Palgrave Macmillan.

Coe, Kevin, Kate Kenski, and Stephen A Rains. 2014. "Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments." *Journal of Communication* 64 (4): 658–679.

Combs, Aidan, Graham Tierney, Brian Guay, Friedolin Merhout, Christopher A. Bail, D. Sunshine Hillygus, and Alexander Volfovsky. 2023. "Reducing political polarization in the United States with a mobile chat platform." *Nature Human Behaviour* 7 (9): 1454–1461.

Connors, Elizabeth C, and Christopher Howell. n.d. ""You Need to Calm Down": Heated Tones Drive Aversion to Political Discussions." *Working Paper,* http://www.elizabethchaseconnors.com/images/Tone.pdf.

Druckman, James N, SR Gubitz, Ashley M Lloyd, and Matthew S Levendusky. 2019. "How Incivility on Partisan Media (De)Polarizes the Electorate." *The Journal of Politics* 81 (1): 291–295.

Eveland, William P., Jr. 2004. "The Effect of Political Discussion in Producing Informed Citizens: The Roles of Information, Motivation, and Elaboration." *Political Communication,* 21 (2): 177–193.

Fishkin, James, Alice Siu, Larry Diamond, and Norman Bradburn. 2021. "Is Deliberation an Antidote to Extreme Partisan Polarization? Reflections on "America in One Room"." *American Political Science Review* 115 (4): 1464–1481.

Forgette, Richard, and Jonathan S Morris. 2006. "High-Conflict Television News and Public Opinion." *Political Research Quarterly* 59 (3): 447–456.

Frimer, Jeremy A, Harinder Aujla, Matthew Feinberg, Linda J Skitka, Karl Aquino, Johannes C Eichstaedt, and Robb Willer. 2023. "Incivility Is Rising Among American Politicians on Twitter." *Social Psychological and Personality Science* 14 (2): 259–269.

Frimer, Jeremy A, and Linda J Skitka. 2018. "The Montagu Principle: Incivility Decreases Politicians' Public Approval, Even With Their Political Base." *Journal of Personality and Social Psychology* 115 (5): 845–866.

Gao, Yujia, Wenna Qin, Aniruddha Murali, Christopher Eckart, Xuhui Zhou, Jacob Daniel Beel, Yi-Chia Wang, and Diyi Yang. 2024. "A Crisis of Civility? Modeling Incivility and Its Effects in Political Discourse Online." In *Proceedings of the International AAAI Conference on Web and Social Media,* 18:408–421.

Gerber, Alan S, and Donald P Green. 2012. *Field Experiments: Design, Analysis, and Interpretation.* New York, NY: W.W. Norton.

Gervais, Bryan 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.

———. 2019. "Rousing the Partisan Combatant: Elite Incivility, Anger, and Antideliberative Attitudes." *Political Psychology* 40 (3): 637–655.

Gibson, James L, and Joseph L Sutherland. 2023. "Keeping Your Mouth Shut: Spiraling Self-Censorship in the United States." *Political Science Quarterly* 138 (3): 361–376.

Goel, Natasha, and Eric Merkley. 2025. "The Nature of Online Talk: Incivility of Opposing Views and Affective Polarization." *Political Behavior* 47:1483–1500.

Goyanes, Manuel, Porismita Borah, and Homero Gil De Zúñiga. 2021. "Social media filtering and democracy: Effects of social media news use and uncivil political discussions on social media unfriending." *Computers in Human Behavior* 120:106759.

Guess, Andrew M, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, et al. 2023. "How do social media feed algorithms affect attitudes and behavior in an election campaign?" *Science* 381:398–404.

Hobolt, Sara B, Katharina Lawall, and James Tilley. 2023. "The Polarizing Effect of Partisan Echo Chambers." *American Political Science Review* 118 (3): 1464–1479.

Hopp, Toby, Chris J Vargo, Lucas Dixon, and Nithum Thain. 2020. "Correlating Self-Report and Trace Data Measures of Incivility: A Proof of Concept." *Social Science Computer Review* 38 (5): 584–599.

Hutto, Clayton, and Eric Gilbert. 2014. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." *Proceedings of the International AAAI Conference on Web and Social Media* 8 (1): 216–225.

Hwang, Hyunseo, Youngju Kim, and Catherine U Huh. 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, Shanto, and Masha Krupenkin. 2018. "The Strengthening of Partisan Affect." *Political Psychology* 39 (S1): 201–218.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. 2012. "Affect, Not Ideology: A Social Identity Perspective on Polarization." *Public Opinion Quarterly* 76 (3): 405–431.

Kim, Jin Woo, Andrew Guess, Brendan Nyhan, and Jason Reifler. 2021. "The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity." *Journal of Communication* 71 (6): 922–946.

Kim, Yonghwan, and Youngju Kim. 2019. "Incivility on Facebook and political polarization: The mediating role of seeking further comments and negative emotion." *Computers in Human Behavior* 99:219–227.

Kooti, Farshad, Karthik Subbian, Winter Mason, Lada Adamic, and Kristina Lerman. 2017. "Understanding Short-term Changes in Online Activity Sessions." In *Proceedings of the 26th International Conference on World Wide Web Companion,* 555–563.

Kumar, Deepak, Jeff Hancock, Kurt Thomas, and Zakir Durumeric. 2023. "Understanding the Behaviors of Toxic Accounts on Reddit." In *Proceedings of the ACM Web Conference 2023,* 2797–2807.

Lapidot-Lefler, Noam, and Azy Barak. 2012. "Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition." *Computers in Human Behavior* 28 (2): 434–443.

Leiter, Michael P., Emily Peck, and Stephanie Gumuchian. 2015. "Workplace Incivility and its Implications for Well-Being." In *Mistreatment in Organizations.* Bingley, England: Emerald, June.

Levendusky, Matthew S, James N Druckman, and Audrey McLain. 2016. "How group discussions create strong attitudes and strong partisans." *Research & Politics* 3 (2): 2053168016645137.

Levendusky, Matthew S, and Neil Malhotra. 2016. "(Mis)perceptions of Partisan Polarization in the American Public." *Public Opinion Quarterly* 80 (S1): 378–391.

Levendusky, Matthew S, and Dominik A Stecula. 2021. *We Need to Talk: How Cross-Party Dialogue Reduces Affective Polarization.* Cambridge, UK: Cambridge University Press.

Lohmann, Sophie, and Emilio Zagheni. 2023. "Diversity of social media use: Self-selection explains associations between using many platforms and well-being." *PLOS Digital Health* 2 (7): e0000292.

Lu, Shuning, Hai Liang, and Gina M Masullo. 2023. "Selective Avoidance: Understanding How Position and Proportion of Online Incivility Influence News Engagement." *Communication Research* 50 (4): 387–409.

Mamakos, Michalis, and Eli J Finkel. 2023. "The social media discourse of engaged partisans is toxic even when politics are irrelevant." *PNAS nexus* 2 (10): pgad325.

Masullo Chen, Gina, and Shuning Lu. 2017. "Online Political Discourse: Exploring Differences in Effects of Civil and Uncivil Disagreement in News Website Comments." *Journal of Broadcasting & Electronic Media* 61 (1): 108–125.

Muddiman, Ashley. 2017. "Personal and Public Levels of Political Incivility." *International Journal of Communication* 11:3182–3202.

Munger, Kevin. 2021. "Don't @ Me: Experimentally Reducing Partisan Incivility on Twitter." *Journal of Experimental Political Science* 8 (2): 102–116.

Mutz, Diana C. 2006. *Hearing the Other Side: Deliberative Versus Participatory Democracy.* Cambridge, UK: Cambridge University Press.

———. 2007. "Effects of "In-Your-Face" Television Discourse on Perceptions of a Legitimate Opposition." *American Political Science Review* 101 (4): 621–635.

———. 2015. *In-Your-Face Politics: The Consequences of Uncivil Media.* Princeton, NJ: Princeton University Press.

Mutz, Diana C, and Byron Reeves. 2005. "The New Videomalaise: Effects of Televised Incivility on Political Trust." *American Political Science Review* 99 (1): 1–15.

Myers, C Daniel. 2022. "The Dynamics of Social Identity: Evidence from Deliberating Groups." *Political Psychology* 43 (2): 237–254.

Noelle-Neumann, Elisabeth. 1974. "The Spiral of Silence a Theory of Public Opinion." *Journal of Communication* 24 (2): 43–51.

Nugier, Armelle, Paula M Niedenthal, Markus Brauer, and Peggy Chekroun. 2007. "Moral and angry emotions provoked by informal social control." *Cognition and Emotion* 21 (8): 1699–1720.

Ognyanova, Katherine, David Lazer, Ronald E Robertson, and Christo Wilson. 2020. "Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power." *Harvard Kennedy School Misinformation Review.*

Oz, Mustafa. 2023. "Social media incivility and engagement: The role of experiences, perceptions, and expectations." *First Monday* 28 (11).

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

Patel, Sarah E, and Matthew Chrisman. 2020. "Incivility through the continuum of nursing: A concept analysis." *Nursing Forum* 55 (2): 267–274.

Rathje, Steve, Jay J Van Bavel, and Sander Van Der Linden. 2021. "Out-group animosity drives engagement on social media." *Proceedings of the National Academy of Sciences* 118 (26): e2024292118.

Robertson, Claire E, Kareena S Del Rosario, and Jay J Van Bavel. 2024. "Inside the funhouse mirror factory: How social media distorts perceptions of norms." *Current Opinion in Psychology* 60:101918.

Rossiter, Erin L, and Taylor N Carlson. 2024. "Cross-Partisan Conversation Reduced Affective Polarization for Republicans and Democrats Even after the Contentious 2020 Election." *The Journal of Politics* 86 (4): 1608–1612.

Settle, Jaime E. 2024. "The promise and peril of interpersonal political communication." *Political Psychology* 46 (S1): 167–212.

Shmargad, Yotam, Kevin Coe, Kate Kenski, and Stephen A Rains. 2022. "Social Norms and the Dynamics of Online Incivility." *Social Science Computer Review* 40 (3): 717–735.

Skoric, Marko M, Qinfeng Zhu, Karolina Koc-Michalska, Shelley Boulianne, and Bruce Bimber. 2022. "Selective Avoidance on Social Media: A Comparative Study of Western Democracies." *Social Science Computer Review* 40 (5): 1241–1258.

Skytte, Rasmus. 2021. "Dimensions of Elite Partisan Polarization: Disentangling the Effects of Incivility and Issue Polarization." *British Journal of Political Science* 51 (4): 1457–1475.

———. 2022. "Degrees of Disrespect: How Only Extreme and Rare Incivility Alienates the Base." *The Journal of Politics* 84 (3): 1746–1759.

Stryker, Robin, Bethany Anne Conway, and J Taylor Danielson. 2016. "What is political incivility?" *Communication monographs* 83 (4): 535–556.

Suhay, Elizabeth, Emily Bello-Pardo, and Brianna Maurer. 2018. "The Polarizing Effects of Online Partisan Criticism: Evidence from Two Experiments." *The International Journal of Press/Politics* 23 (1): 95–115.

Suler, John. 2004. "The Online Disinhibition Effect." *CyberPsychology & Behavior* 7 (3): 321–326.

Sun, Qiusi, Magdalena Wojcieszak, and Sam Davidson. 2021. "Over-Time Trends in Incivility on Social Media: Evidence From Political, Non-Political, and Mixed Sub-Reddits Over Eleven Years." *Frontiers in Political Science* 3:741605.

Sydnor, Emily. 2019. *Disrespectful Democracy: The Psychology of Political Incivility.* New York, NY: Columbia University Press.

———. 2020. "Platforms for Incivility: Examining Perceptions Across Different Media Formats." In *Studying Politics Across Media,* 97–116. Routledge.

van Loon, Austin, Srikar Katta, Christopher A. Bail, D. Sunshine Hillygus, and Alexander Volfovsky. 2025. "Designing Social Media to Promote Productive Political Dialogue on a New Research Platform." *OSF Preprints,* https://doi.org/10.31235/osf.io/dngcj_v2.

Van't Riet, Jonathan, and Aart Van Stekelenburg. 2022. "The Effects of Political Incivility on Political Trust and Political Participation: A Meta-Analysis of Experimental Research." *Human Communication Research* 48 (2): 203–229.

Wang, Meredith Y, and David E Silva. 2018. "A slap or a jab: An experiment on viewing uncivil political discussions on facebook." *Computers in Human Behavior* 81:73–83.

Wojcieszak, Magdalena. 2011. "Deliberation and Attitude Polarization." *Journal of Communication* 61 (4): 596–617.

Wojcieszak, Magdalena, and Benjamin R Warner. 2020. "Can Interparty Contact Reduce Affective Polarization? A Systematic Test of Different Forms of Intergroup Contact." *Political Communication* 37 (6): 789–811.

Wulczyn, Ellery, Nithum Thain, and Lucas Dixon. 2017. "Ex Machina: Personal Attacks Seen at Scale." In *Proceedings of the 26th International Conference on World Wide Web,* 1391–1399.

Yang, JungHwan, Hernando Rojas, Magdalena Wojcieszak, Toril Aalberg, Sharon Coen, James Curran, Kaori Hayashi, Shanto Iyengar, Paul K Jones, Gianpietro Mazzoleni, 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.

Yeomans, Michael, Alejandro Kantor, and Dustin Tingley. 2018. "The politeness Package: Detecting Politeness in Natural Language." *R Journal* 10 (2): 489–502.

Yeomans, Michael, Julia Minson, Hanne Collins, Frances Chen, and Francesca Gino. 2020. "Conversational receptiveness: Improving engagement with opposing views." *Organizational Behavior and Human Decision Processes* 160:131–148.

# Online Supplementary Information

## "The Causal Effects of Political Incivility in Social Media Discussions"

**Table of Contents**


**A Study Details** **2**

A.1 Population and Recruitment . . . 2

A.2 Missing Data . . . 2

A.3 Respondent Experience . . . 2

A.4 Synthetic User Prompts, Tags, and Profiles . . . 3

A.5 Newsfeed Design and Content . . . 5

A.6 Pre-Testing of Study Materials . . . 6

A.6.1 Text-Based Measures . . . 7

A.6.2 Human Evaluations . . . 8

A.7 Data Quality . . . 9

A.8 AAPOR Disclosures . . . 10

A.9 Questionnaires . . . 10

A.9.1 Screening Survey . . . 10

A.9.2 Pre-treatment Survey . . . 11

A.9.3 Post-treatment Survey . . . 12

**B Descriptive Statistics** **13**

**C Attrition Analysis** **19**

**D Supplemental Tables, Figures, and Analyses** **21**

# SI A Study Details

## A.1 Population and Recruitment

Participants in our study were recruited from Prolific (https://www.prolific.com/), an online, opt-in platform where interested individuals can participate in a variety of tasks (including research studies) in exchange for payment. Prolific maintains its own pool of respondents who have been vetted for quality assurance purposes (read more here). Our final analytical sample consisted of $N = 1,461$ who were assigned to one of the two experimental conditions, failed no more than one data quality check, and successfully completed all stages of the study including: eligibility survey, pre-treatment survey, app download, newsfeed experience, and post-treatment survey. The recruitment message accompanying our Prolific task is given below:

> *We would like to know your opinions about a new social media platform. Spend 30 minutes trying out the platform and sharing your thoughts! Click the link below to be taken to an intake survey. Once complete, you will be given instructions on how to get access to the platform through the "Spark Social" testing environment. You will need to download Spark Social onto your mobile device to complete this task.*
>
> *To be eligible for this study, you must:*
>
> 1. *Be a US resident*
> 2. *Be at least 18 years of age*
> 3. *Own an Apple iPhone or an Android device*
> 4. *Be willing to download "Spark Social" onto your smart phone.*

## A.2 Missing Data

Participants who attritted from the study had missing data for survey questions asked after attrition. Our approach to handling these missing data in our attrition analyses are described in SI C. Our final analytical sample, which excluded all attritted participants, had almost no item non-response (see Tables S2 and S3). Note, however, that our in-party and out-party thermometer ratings had 144 missing values for self-identified 'pure independents' who do not have an in-/out-party to reference.

## A.3 Respondent Experience

A flowchart of the respondent experience is provided in Figure S1. Starting at the top, respondents come the eligibility survey on a device of their choosing. Those who fail to meet eligibility criteria or do not pass attention checks are excluded at this point. Those who meet these criteria are then given a participant code along with instructions for downloading the Spark Social mobile app through either the Apple App Store or the Google Play Store. Once participants download and open the app, they enter their invite code and complete an embedded pre-treatment Qualtrics survey, after which they are given instructions on using the newsfeed in a short on-boarding process. Those instructions are given below.

> *Thank you for your responses. We will now move to the next stage where you will help test our new social media platform. On our platform, we're looking to build online communities of people from across the country. To kickstart your experience, you have been randomly selected to join an ongoing conversation on energy, climate, and the environment. Continue to the next page to learn how to use our platform so you*

*may engage in the conversation.*

*Below is a picture of the newsfeed which displays new posts from other users in your conservation. Below each post in your newsfeed, you'll find a set of buttons that enable various interactions with the content. With these buttons, you can "Like" (thumbs-up), "Dislike" (thumbs-down), and "Comment" on (speech bubble) content you see in your newsfeed.*

*Tap the box at the top of the newsfeed to author your own original post and share your thoughts with others in your conversation!*

*When you enter the newsfeed, feel free to look around and interact as you wish. We want to ensure we get your honest feedback about the platform and your experience. Finally, it is important to understand that you will have exactly 12 minutes to experience the newsfeed. After that allotted time, you will be automatically redirected to a final survey, during which you will be asked about your experience and opinions.*

Next, respondents enter the civil or uncivil feed to which they were assigned and remain for 12 minutes, interacting as they wish. When time is up, participants are automatically re-routed to a post-treatment Qualtrics survey within the app, receiving a compensation code to enter on Prolific at the end of the survey. More on attrition at these various stages of the process is given in SI C.

## A.4 Synthetic User Prompts, Tags, and Profiles

The Social Media Accelerator platform used in this study features synthetic users powered by OpenAI's GPT-4 large language model which are capable of dynamically interacting with research participants through posts, comments, engagements, and 'following' behavior. This study makes specific use of the dynamic commenting feature to create environments where participants can observe or (if they wish) engage in political discussions with the synthetic users. The behavior of the synthetic users is guided by the prompt shown below. This prompt uses the "RTF" or "Role-Task-Format" framework which starts by defining the persona or role that GPT is to play (role), then specifies the function that GPT is to perform (task), and finally offers directions on structuring the model output (format).

**Comment Prompt:**

*Here are some tags: {tags}. Item 1 is your name. Item 2 is your personality. Item 3 is your bio. Item 4 is your political orientation and opinions. Write a short response between 80 and 160 characters in the style of your personality to either join or continue the online conversation below. [ONLY refer to your bio or political orientation IF IT IS RELEVANT TO THE CONVERSATION. DO NOT tell me your name or your partisanship. DO NOT use hashtags or emojis. Use informal language]*

*[POST AND ANY ASSOCIATED COMMENTS PASTED HERE]*

While the commenting behavior of all synthetic users is guided by the same prompt shown above, each user has a unique set of 'tags' to describe their personality and political opinions. These tags are automatically piped into the prompt once a user is selected to

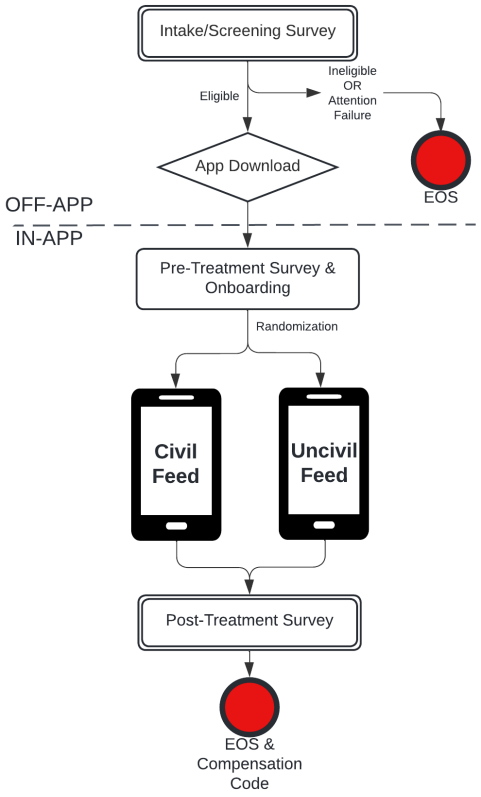



Figure S1: Flowchart of the study procedure

Note: EOS = End of survey. Eligibility criteria include consenting to participation, being 18 years of age or older, residing within the United States, owning an Apple iPhone or Android device, being willing to download the Spark Social app, and passing attention checks.

produce a comment. As mentioned in the main text, one of our primary experimental manipulations occurs in the synthetic user tags, where we describe them with traits more typical of civil or uncivil discussants. An example of the tags used in this study are given below.

**Example Tags:**

- **Civil Republican:** {Sammy11; a reserved person; a stock trader; a Republican who supports traditional American energy like oil and coal, and isn't sold on green energy solutions}
- **Civil Democrat:** {Siriusxo; a frugal person; a sales supervisor; a Democrat who supports carbon taxes and opposes the local agricultural industry}
- **Uncivil Republican:** {Cjallday: a very disrespectful and toxic person; a marketing specialist; a Republican who supports environmental deregulation and opposes government subsidies for electric vehicles}
- **Uncivil Democrat:** {ItzErin; a very impolite and abrasive person; a regulatory professional; a Democrat who supports environmental regulation, and opposes fracking and drilling}

There are two ways in which comments are generated by synthetic users. Both of these procedures are described in detail below:

### Fixed-Interval Comment Generation Procedure:

1. Every 65 seconds, generate a comment using the 'event-based' comment generation procedure below

### Event-Based Comment Generation Procedure:

1. Randomly select a post in the newsfeed
    - If post has no comments, all bots have equal probability of being selected to comment.
    - If post author is a study participant and post has been commented upon by a bot, the same bot has a 75% chance of commenting again, while 25% chance is given to a new bot joining the discussion.
    - If post author is a bot and a bot has already commented, there is a 50% chance of that bot commenting again, a 35% chance of it being the bot author, and 15% chance a different bot is selected.
2. Generate bot prompt and make API call
3. Get return call from API with result
4. Check result and cancel if (1) contains "bot" content (e.g., "as an AI language model..."), (2) toxicity is greater than 0.95, or (3) result is duplicate from same study and post.
5. If passing, store result in database and post comment.
6. After 20 seconds, bots are triggered to reply with 65% probability following the same procedure (EVENT-BASED ONLY).
7. If enabled at study-level, notification is sent to participant if comment is on post authored by participant or they previously commented on the post (EVENT-BASED ONLY).

## A.5 Newsfeed Design and Content

The newsfeeds of both the civil and uncivil treatment arms were pre-populated with 12 posts and 12 comments when participants initially enter the newsfeed, all of which was manually crafted by the research eam. A schematic of the newsfeed is provided in Table S1 showing the name and partisanship of each posting "user"; the type, topic, and timing of the post; the number of likes and dislikes; and the type and partisanship of each commenter. Note here that a "majority"-type post or comment is one that is civil in the civil newsfeed or uncivil in the uncivil newsfeed. The inverse is true for "minority"-type content (i.e., civil in uncivil newsfeed; uncivil in civil newsfeed). Real examples of this content from the civil and uncivil newsfeeds are provided below.

### Sample Content - Civil Newsfeed

1. Sammy11 (R-Civil): "The issue with windmills and solar panels is that they are weather-dependent. They don't seem all that reliable and I'm not confident we've thought that thru yet."
2. Siriusxo (D-Civil): "My family had to evacuate twice last year because of wildfires and my cousin in Florida had her house flooded in a hurricane. We really need to take action to address these environmental issues."
3. Mwilson200 (R-Civil): "As a child of Appalachia myself, coal has been part of my life and my families' lives for decades. I don't think this new green energy stuff has been good for the industry or my people."

Table S1: Schematic of Spark Social newsfeed

| Name | Party | Post type | Post topic | Timing* | Likes | Dislikes | Comments | Commenter party-type |
|---|---|---|---|---|---|---|---|---|
| Taylor017 | D | Majority | Action on climate change | +10 | 0 | 0 | 0 | |
| Real_ET | R | Majority | Effect of Dem. environmental policy | +8 | 0 | 0 | 0 | |
| Lazyart | D | Majority | Criticism of Rep. environmental policy | +4 | 0 | 0 | 0 | |
| TheOneTwo | R | Majority | Opening of local oil refinery | +2 | 0 | 0 | 0 | |
| Sammy111 | R | Majority | Reliability of wind and solar | -2 | 0 | 3 | 0 | |
| Siriusxo | D | Majority | Extreme weather | -8 | 2 | 3 | 0 | |
| Mwsilon200 | R | Majority | Loss of coal industry | -15 | 3 | 3 | 3 | R-Maj, D-Min, R-Maj |
| Lazyart | D | Majority | Electric vehicles | -22 | 10 | 5 | 3 | D-Maj, R-Min, D-Maj |
| AloofGoose | - | Filler (Off-topic) | Chicken wings | -31 | 10 | 2 | 0 | |
| Cjallday | R | Minority | Food access in rural areas | -36 | 3 | 9 | 1 | D-Maj |
| Spark Social | - | Filler (Bonus) | Bonus code | -38 | 15 | 0 | 0 | |
| ItsErin | D | Majority | Carbon tax on corporations | -42 | 10 | 5 | 2 | R-Maj, D-Maj |
| AloofGoose | - | Filler (Off-topic) | Baseball | -45 | 10 | 0 | 0 | |
| Real_ET | R | Majority | Livestock, methane, and hamburgers | -46 | 8 | 2 | 2 | D-Maj, R-Maj |
| ThaGemini2 | D | Minority | Drought and rising temperatures | -60 | 3 | 12 | 1 | R-Maj |
| Spark Social | - | Filler (Bonus) | Bonus code | -65 | 18 | 0 | 0 | |

*Note:* R = Republican Party; D = Democratic Party
Maj = Majority Condition = Civil content in Civil Condition, Uncivil content in Uncivil condition
Min = Minority Condition = Uncivil content in Civil Condition, Civil content in Uncivil condition
*Timing refers to the timestamp of each post relative to when a participant enters the feed. Negative values indicate posts that were present upon entry; positive values indicate posts that appeared after entry.

(a) Sammy11 (R-Civil): "Agreed, coal forms the foundation of many communities and losing that can be hard on folks. "
(b) ThaGemini2 (D-Uncivil): "Maybe just. . . adapt? Coal is awful for the environment and people's health. Don't be complicit in poisoning others!"
(c) Mwilson200 (R-Civil): "I get where you're coming from, but let's not forget the jobs and economies built around coal. We should have some empathy for those being affected by the loss of industry."

### Sample Content - Uncivil Newsfeed

1. Sammy112 (R-Uncivil): "Wind mills and solar panels only ever work half the time, if that!! It's an idiotic solution and yet we're running blindly with it."
2. SiriusXO (D-Uncivil): "My family had to evacuate TWICE last year because of wildfires AND my cousin in Florida had her house flooded in a hurricane. I don't want to hear a single word from the idiots who think the solution is for government to ignore our environmental issues."
3. Mwilson300 (R-Uncivil): "As a child of Appalachia myself, coal has been part of my life and my families' lives for generations. Now these green energy scammers are throttling the industry and taking my people to hell with em!!!"
   (a) Sammy12 (R-Uncivil): "I feel ya, these people attacking coal don't realize that they are tearing apart the foundations of many communities! Its honestly heartless."
   (b) ThaGemini3 (D-Civil): "I think there's a case to be made that industries should change with the times. Plus renewables are a lot better from a health perspective compared to coal."
   (c) Mwilson300 (R-Uncivil): 'Maybe we can change eventually but for right now they are killing the jobs and the economies built by coal!! These people have ZERO empathy."

## A.6 Pre-Testing of Study Materials

Before launching the study, we pretested our study materials (i.e., researcher-generated posts and comments; GPT-generated comments) to ensure that we have sufficiently manipulated the level of incivility in this content. The posts and comments are the researcher-generated content that pre-populates or is scheduled to enter the newsfeed. The comments

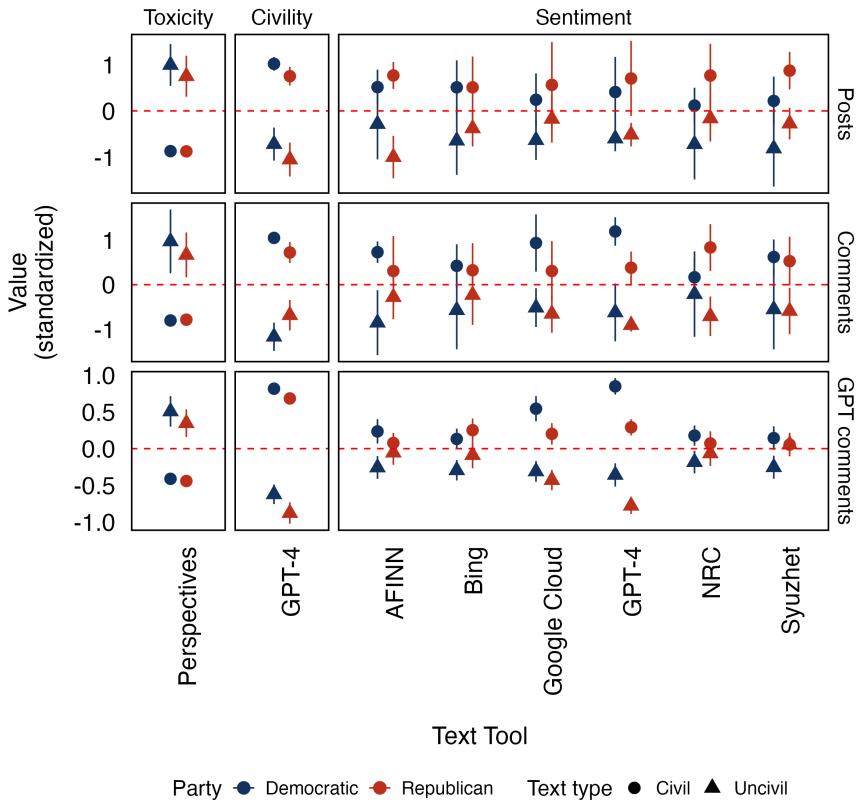



Figure S2: Measuring text features of research-generated posts, comments, and GPT comments

from our GPT-powered synthetic users were created by having each of our 10 users respond to each of the 24 researcher-generated posts with either a "civil" or "uncivil" persona. This allows us to determine whether our manipulation of a synthetic user's civility can be detected while holding all other features of the user constant, including the content to which its responding. We pretest our materials using a set of text-based measures of toxicity, civility, and sentiment (A.6.1), and human evaluations from participants on CloudResearch's "Connect" platform (A.6.2).

### A.6.1 Text-Based Measures

Figure S2 plots the standardized values of various text-based measures of the researcher-generated posts (top panel) and comments(middle panel), as well as GPT-generated comments (bottom), broken down by civility (shape) and partisanship (color) of the posting "user." There is clear variation by measure, but the civil and uncivil content is generally distinguishable in all forms of content, and in roughly equal amounts across party lines.

### A.6.2 Human Evaluations

We also assessed the efficacy of the civility manipulation in our researcher- and GPT-generated content using human evaluations from participants on CloudResearch's "Connect" platform. We do this with two exercises, the first asking participants to rate the civility or incivility of a single piece of content on a 7-point scale from "Very uncivil" to "Very civil" (absolute rating) and the second asking participants to choose the more uncivil of two pieces of content (relative rating). Respondents were asked to rate two pieces (or pairs) of each type of content (i.e., posts, comments, and GPT-generated comments).

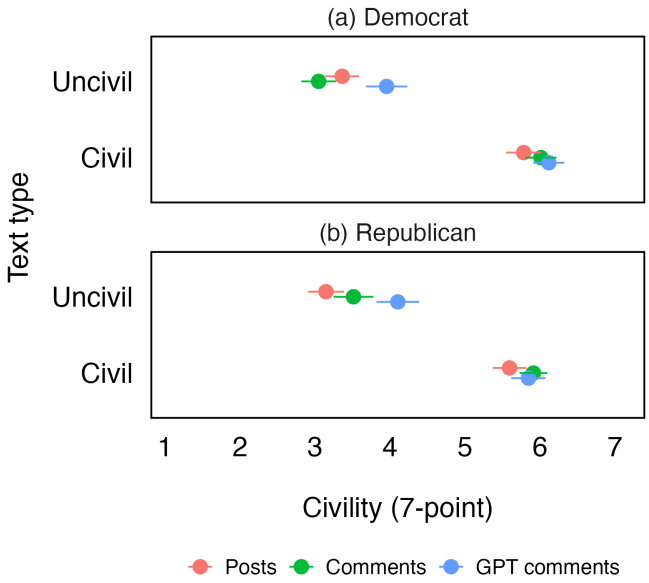



Figure S3: Mean civility ratings of newsfeed content by content type and partisanship of contributing profile (Connect pretest)

Figure S3 shows the mean civility ratings across the types of content, as well as the researcher-assigned civility and partisanship of the contributing user. Here we see that the all forms of civil content from both Democratic and Republican users are rated as moderately civil (∼5.75). Similarly, all forms of uncivil content from both Democratic and Republican profiles are rated as mildly uncivil (∼3.75), though the GPT-generated comments are rated as slightly more mild than the researcher-generated content. Overall, these findings give us confidence that we have successfully manipulated civility in our content, and to roughly equal extents across the Democratic and Republican content.

Figure S4 shows the proportion of respondents that correctly or incorrectly selected the more uncivil of two pieces of content. Here we see that roughly 90% of the time respondents correctly labeled the uncivil content as uncivil, and at roughly equal rates across each type of content and across Democratic and Republican profiles. These results provide further assurance that respondents can detect our incivility manipulation across the various types of content shown in the feed.

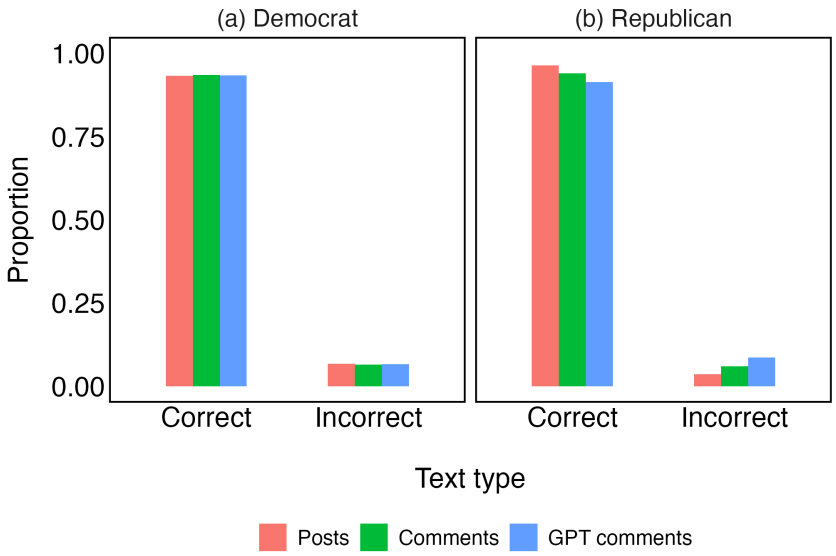



Figure S4: Share correctly and incorrectly identifying uncivil content by type and partisanship of contributing profile (Connect pretest)

## A.7 Data Quality

To ensure the data we collect are fit for use, we included a series of data quality measures throughout our study. First, we include a red herring item and an attention check to immediately screen inattentive or potentially fraudulent respondents from the study. The red herring item is placed in the screening survey and asks about respondents' use of a fake social platform called "ChatterChain," which screened out 39 who said they used this fake platform "some" or "a lot." The attention check also appeared in the screening survey and asked respondents to identify a number shown in a box; this task was passed by all.

We also look for indicators of poor data quality among those completing all stages of the study by creating flags for excessive speed, missing or mismatched location data (state and ZIP), and non-response/non-sequitur to an open-ended question about the download process. We define excessive speed as completing the pre-treatment survey in a time less than one-third the median time of all respondents; five respondents received a speeding flag.[35] Eight additional respondents received flags for missing ZIP code data, providing a ZIP code that does not match a US state, or providing a ZIP that does not match the respondents reported state of residence. Five more respondents received flags for providing no response or a non-sequitur to an open-ended question asking respondents to describe their experience with the download process (pre-treatment survey). Wording for all data quality items are given in SI A.9. Overall, the vast majority of respondents (99.0%) received no data quality flags, several received only one flag (14; 1.0%), and only one respondent received two or more

[35]Time spent completing the pre-treatment survey is calculated as the end time minus the start time minus the timing for the last item in the survey (i.e., final page of on-boarding). This calculation was used to ensure timings were not influenced by fast respondents sitting on the last page of the pre-treatment survey to make their timings appear less suspicious.

flags. This one respondents is removed from all analyses as specified in our pre-registration.

## A.8 AAPOR Disclosures

- **Data collection strategy**: a survey experiment conducted with the Qualtrics survey software and the "Social Media Accelerator" research platform with participants recruited from Prolific (https://www.prolific.com/).
- **Who Sponsored the Research and Who Conducted It:** This research was funded in part by the John Templeton Foundation (Award No. 62656), NSF CAREER (Award No. DMS-2046880), The Carnegie Foundation, a Facebook Foundational Research Award, and Duke University. Research conducted by members of the Duke Polarization Lab at Duke University.
- **Measurement Tools/Instruments** Survey items used in the screening, pre-treatment, and post-treatment surveys provided in SI A.9. Wording using in recruitment and participant instructions provided in SI A.3.
- **Population Under Study** Adults 18+ residing in the U.S.
- **Method Used to Generate and Recruit the Sample** We use a non-probability, opt-in sample from the Prolific recruitment platform. Participants must be 18 years of age or older, reside within the US, and own an Android or Apple phone capable of downloading the "Spark Social" mobile application. Partisanship quotas were included to ensure our sample contained adequate numbers of Democrats, independents, and Republicans with varying strengths of identification (33% Republican/33% Democratic/34% Independent).
- **Method(s) and Mode(s) of Data Collection** Participants were recruited from the Prolific platform via the web and completed the screening survey in Qualtrics through a web-capable device of their choosing. Participants received instructions at the end of the screening survey on how to download the "Spark Social" app to their mobile device via the Apple App Store or Google Play Store. Within the mobile app, participants completed a pre-treatment Qualtrics survey before participants in the newsfeed experience for a total of 12 minutes and performing a final, post-treatment Qualtrics survey. All stages of the study were conducted in English.
- **Dates of Data Collection** Tuesday, June 18, 2024 - Monday, June 24, 2024
- **Sample Sizes**: A total of 1,461 complete submissions passing our data quality checks as specified in pre-registration. See SI C for more on attrition.
- **How the Data Were Weighted**: Data in this study were not weighted.
- **Statement of Limitations**: We acknowledge that all forms of public opinion research, including the present study, face limitations and unmeasured forms of error.

## A.9 Questionnaires

### A.9.1 Screening Survey

1. How much, if at all, have you used the following social media platforms in the past week? [Facebook/X (formerly "Twitter")/YouTube/TikTok/Instagram/Truth Social/Threads/Snapchat/Reddit/ChatterChain] [DATA QUALITY/RED HERRING]

2. In which state do you currently reside? [Dropdown with 50 US States, D.C., and Puerto Rico][ELIGIBILITY]
3. What is your age in years? Please enter a whole number [text entry limited to 0-120] [ELIGIBILITY]
4. Is your mobile phone an Apple iPhone, an Android, or something else? [Apple iPhone/Android/ Something Else] [ELIGIBILITY]
5. Are you willing and able to download an app from the [Apple App Store/Google Play Store] to complete this research? You may delete the app after completing the study. [Yes, I am/No, I am NOT] [ELIGIBILITY]
6. What is the highest level of school you have completed or the highest degree you have received? [Less than high school/High school graduate or equivalent (GED)/Some college/Associate's degree/Bachelor's degree/Post-graduate degree]
7. What racial/ethnic group(s) do you consider yourself to be a part of? Select all that apply [Asian/Black or African American/Hispanic/White/Something else]
8. Please select the number shown in the box above. [7/11/3/54/20][DATA QUALITY/ATTENTION CHECK]
9. Do you usually think of yourself as a Democrat, a Republican, an Independent, or what? [Democrat/Republican/Independent/Other]
10. (For Democrats and Republicans) Would you call yourself a strong [Democrat/Republican] or a not very strong [Democrat/Republican]? [Strong PARTY, Not very strong PARTY]
11. (For Independent and Other) Do you think of yourself as closer to the Republican Party or to the Democratic Party? [Closer to Republican Party/Neither/Closer to Democratic Party]
12. Where would you place yourself on this scale? [Extremely liberal/Liberal/Somewhat liberal/Moderate or middle of the road/Somewhat conservative/Conservative/Extremely conservative]

### A.9.2 Pre-treatment Survey

1. How often, if at all, do you pay attention to what's going on in government and politics? [Almost never/Once in a while/About half the time/Most of the time/All of the time]
2. Those at Point 1 want the government in Washington to play a much smaller role in addressing climate change while those at Point 7 want government playing a much larger role in addressing climate change. Others have attitudes in between at Points 2-6. Where would you place [Democratic/Republican/yourself] voters on the scale? [1 Government play a much smaller role in addressing climate change/2/3/4/5/6/7 Government play a much larger role in addressing climate change][PRE-POST]
3. Now we'd like you to rate some groups on a scale from 0 (most cold/negative) to 100 (most warm/positive). You would rate the group at 50 if you feel neither negative nor positive toward them. How would you rate [Democratic/Republican] voters? Enter a whole number. [0-100][PRE-POST]
4. How often, if ever, can you trust the government in Washington to do what is right? [Never/Some of the time/About half the time/Most of the time/Always][PRE-POST]
5. How satisfied are you, if at all, with the way democracy is working in the United States? [Extremely satisfied/Satisfied/Somewhat unsatisfied/Neither satisfied nor unsatisfied/Unsatisfied/Extremely unsatisfied][PRE-POST]
6. Finally, please enter the 5-digit ZIP code where you reside. Please enter a whole number [text entry][DATA QUALITY/ATTENTION CHECK]

### A.9.3 Post-treatment Survey

Note: Measures taken in both the pre- and post-treatment surveys are marked [PRE-POST] in SI A.9.2 above and are not repeated here for brevity.

1. How comfortable, if at all, did you feel sharing your views about political issues on this app? [Not at all comfortable/Slightly comfortable/Moderately comfortable/Very comfortable/Extremely comfortable]
2. During your time on our platform, you may have seen a post that displayed a "super secret password". If you saw this post, what was the password? (If you didn't see this post, continue to the next page.)[text entry]
3. All social media platforms are confronting issues around the presence of bots these days. Do you think you encountered any bots during your time on the platform? [No, I did not encounter any bots/Yes, I encountered one or two bots/Yes, I encountered several bots/Yes, I encountered many bots]
4. Thinking about the discussion you saw on the platform, would you say those discussions were mostly civil or mostly uncivil? [Mostly civil/Mostly uncivil] [MANIPULATION CHECK]
5. Is there anything else we should know about your experience on our platform? For instance, did you experience any harassment or technical glitches? [text entry]

# SI B Descriptive Statistics

Table S2: Summary statistics for categorical variables

| Variable | N | Percent (%) |
|---|---|---|
| Device | 1,461 | |
| ... Android | 600 | 41.1 |
| ... Apple iPhone | 861 | 58.9 |
| Education | 1,461 | |
| ... College | 842 | 57.6 |
| ... High school | 234 | 16.0 |
| ... Less than high school | 9 | 0.6 |
| ... Some college | 376 | 25.7 |
| Gender Identity | 1,461 | |
| ... Man | 482 | 33.0 |
| ... Something else | 23 | 1.6 |
| ... Woman | 956 | 65.4 |
| Partisanship | 1,461 | |
| ... Democratic | 509 | 34.8 |
| ... Independent | 456 | 31.2 |
| ... Other | 20 | 1.4 |
| ... Republican | 476 | 32.6 |
| Race | 1,461 | |
| ... Asian | 58 | 4.0 |
| ... Black | 188 | 12.9 |
| ... Hispanic | 72 | 4.9 |
| ... Mixed/Other | 142 | 9.7 |
| ... White | 1001 | 68.5 |

Note: Table shows summary statistics for categorical demographic and political variables. Final analytical sample size of $N = 1,461$, limited to those who completed all three surveys (screening, pre-treatment, post-treatment), the newsfeed experience, and had $< 2$ data quality flags.

Table S3: Summary statistics for continuous variables

| Variable | N | Mean | Range | SD |
|---|---|---|---|---|
| Age | 1,461 | 36.4 | [18, 78] | 11.7 |
| Comfort sharing political opinions | 1461 | 3.2 | [1, 5] | 1.3 |
| Partisan voter placements | | | | |
| ... Pretreat Democratic | 1,461 | 5.8 | [1, 7] | 1.4 |
| ... Pretreat Republican | 1,461 | 2.7 | [1, 7] | 1.8 |
| ... Posttreat Democratic | 1,459 | 5.9 | [1, 7] | 1.3 |
| ... Posttreat Republican | 1,459 | 2.5 | [1, 7] | 1.7 |
| Partisan voter thermometers | | | | |
| ... Pretreat In-party | 1,317 | 66.8 | [0, 100] | 20.0 |
| ... Posttreat In-party | 1,317 | 66.4 | [0, 100] | 20.6 |
| ... Pretreat Out-party | 1,317 | 29.2 | [0, 100] | 20.8 |
| ... Posttreat Out-party | 1,317 | 29.0 | [0, 100] | 21.1 |
| Trust in Government | | | | |
| ... Pretreat trust | 1,460 | 2.2 | [1, 5] | 0.8 |
| ... Posttreat trust | 1,461 | 2.2 | [1, 4] | 0.8 |
| Satisfaction with democracy | | | | |
| ... Pretreat satisfaction | 1,461 | 2.8 | [1, 7] | 1.5 |
| ... Posttreat satisfaction | 1,461 | 2.9 | [1, 7] | 1.5 |

Note: Table shows summary statistics for continuous demographic and political variables. Final analytical sample size of $N = 1,461$, limited to those who completed all three surveys (screening, pre-treatment, post-treatment), the newsfeed experience, and had $< 2$ data quality flags.

Table S4: Comparing sample to general population and social media users

| Variable | Our sample | ANES full | ANES social media users | ANES political posters |
|---|---|---|---|---|
| Age | 36.4 | 48.4 | 40.0 | 47.0 |
| Gender (%) | | | | |
| ... Man | 33.0 | 47.6 | 46.2 | 48.3 |
| ... Woman | 65.4 | 50.5 | 51.4 | 48.4 |
| ... Something else | 1.6 | 1.9 | 2.4 | 3.3 |
| Partisanship (%) | | | | |
| ... Democratic | 49.7 | 47.7 | 52.0 | 50.6 |
| ... Republican | 40.5 | 45.1 | 41.8 | 44.8 |
| ... Independent | 9.9 | 7.2 | 6.2 | 4.6 |
| Race (%) | | | | |
| ... Asian | 4.0 | 4.8 | 4.5 | 4.5 |
| ... Black | 12.9 | 11.8 | 12.3 | 13.2 |
| ... Hispanic | 4.9 | 14.1 | 15.5 | 14.7 |
| ... Mixed/Other | 9.7 | 5.0 | 5.4 | 6.1 |
| ... White | 68.5 | 64.3 | 62.3 | 61.5 |
| College degree (%) | 57.6 | 45.3 | 51.2 | 44.3 |
| Social media use (1-7) | 5.3 | 2.9 | 5.1 | 3.9 |
| Political interest (1-5) | 3.3 | 3.5 | 3.4 | 3.8 |
| Party thermometers (0-100) | | | | |
| ... In-party | 66.8 | 71.3 | 67.8 | 75.0 |
| ... Out-party | 29.2 | 19.2 | 19.5 | 15.5 |

Note: Table shows summary statistics for categorical demographic and political variables across our original sample, the full 2024 ANES sample, social media users in the ANES, and ANES social media users who report posting online about politics. Data from the 2024 ANES Preliminary Release (August 8, 2025 version). 'ANES full' is the entire 2024 sample; 'ANES social media users' consists of 2024 respondents who mentioned they had visited four or more social media sites in the last year (including Facebook, Twitter, Instagram, Reddit, YouTube, Snapchat, and Tiktok); 'ANES political posters' consists of respondents who reported posting information about political issues or candidates on Facebook or Twitter at least sometimes.

Table S5: Summary statistics for posts and comments

| Statistic | All | Civil | Uncivil | P-Value |
|---|---|---|---|---|
| Posts | | | | |
| Number of posts | 964.0 | 458.0 | 506.0 | – |
| Number of unique posters | 807.0 | 384.0 | 423.0 | – |
| Mean characters per post | 105.0 | 107.1 | 103.1 | 0.5 |
| Mean tokens per post | 19.6 | 20.0 | 19.3 | 0.6 |
| Mean toxicity per post | 0.1 | 0.1 | 0.1 | 0.0 |
| Proportion of respondents making 1+ post | 0.6 | 0.5 | 0.6 | 0.0 |
| Mean total characters per respondent | 69.2 | 66.7 | 71.6 | 0.4 |
| Mean total tokens per respondent | 12.9 | 12.4 | 13.4 | 0.3 |
| Mean number of posts | 0.7 | 0.6 | 0.7 | 0.1 |
| Comments | | | | |
| Number of comments | 4440.0 | 2162.0 | 2278.0 | – |
| Number of unique commenters | 1161.0 | 588.0 | 573.0 | – |
| Mean characters per comment | 79.0 | 83.9 | 74.4 | 0.0 |
| Mean tokens per comment | 14.9 | 15.8 | 14.0 | 0.0 |
| Mean toxicity per comment | 0.1 | 0.0 | 0.1 | 0.0 |
| Proportion of respondents making 1+ comments | 0.8 | 0.8 | 0.8 | 0.5 |
| Mean total characters per respondent | 232.8 | 239.5 | 226.0 | 0.4 |
| Mean total tokens per respondent | 43.8 | 45.1 | 42.6 | 0.3 |
| Mean number of comments | 3.0 | 2.9 | 3.1 | 0.2 |

Note: Table shows summary statistics of posting and commenting behavior in the newsfeed for all respondents, and for those in the civil and uncivil newsfeed, along with a p-value associated with the difference across conditions. P-value is from t-test between the 'civil' and 'uncivil' conditions.

Table S6: Balance table comparing respondents assigned to the civil and uncivil newsfeeds using either $\chi^2$ or Kolmogorov-Smirnov tests.

| Variable | Test | $p$ |
|---|---|---|
| Education | Chi squared | 0.25 |
| Gender | Chi squared | 0.78 |
| Race | Chi squared | 0.43 |
| 3 point Party Identity | Chi squared | 0.91 |
| Age | KS test | 0.69 |
| Pre-treatment in-party thermometer | KS test | 0.28 |
| Pre-treatment out-party thermometer | KS test | 0.44 |
| Pre-treatment 7-point placement (Democratic) | KS test | 1.00 |
| Pre-treatment 7-point placement (Republican) | KS test | 0.44 |
| Pre-treatment political trust | KS test | 0.93 |
| Pre-treatment satisfaction with democracy | KS test | 0.30 |

Note: Table shows $p$-values from balance tests across the civil and uncivil newsfeeds, using either chi squared or KS tests as appropriate
P-value is from t-test between the 'civil' and 'uncivil' conditions

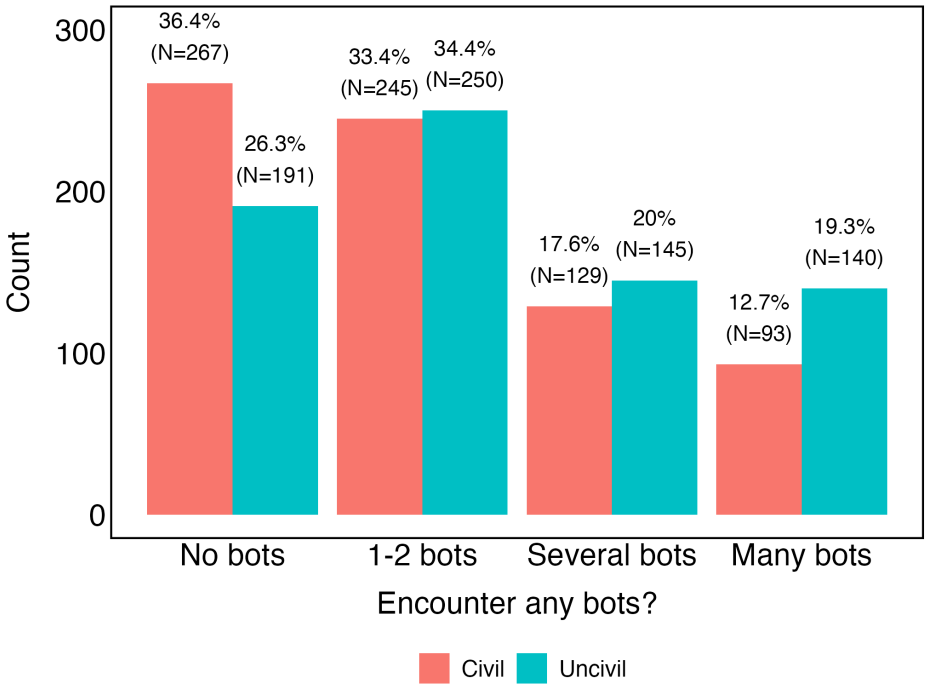



Figure S5: Responses to bot encounters question on post-treatment survey

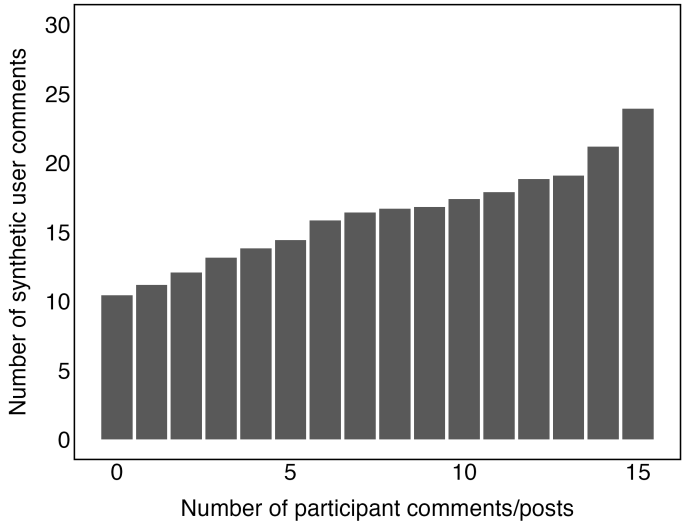
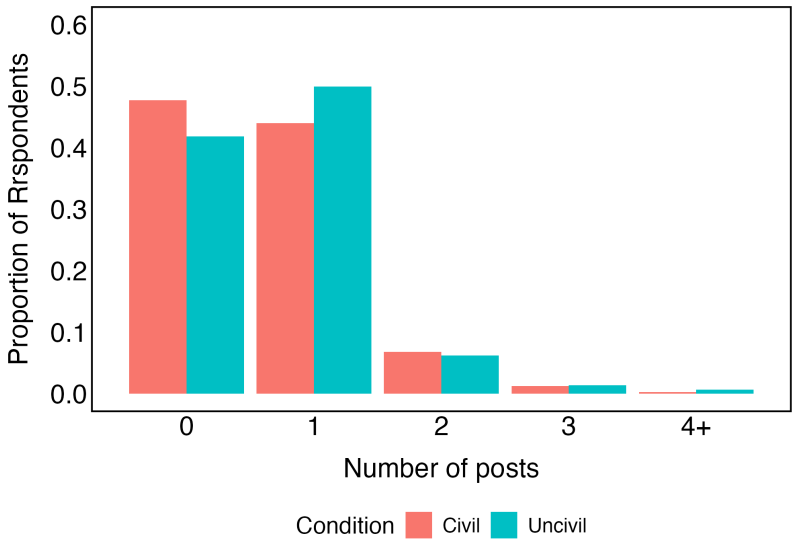



Figure S6: Distribution of synthetic user comments by number of participant posts/comments

(a) Posts




(b) Comments

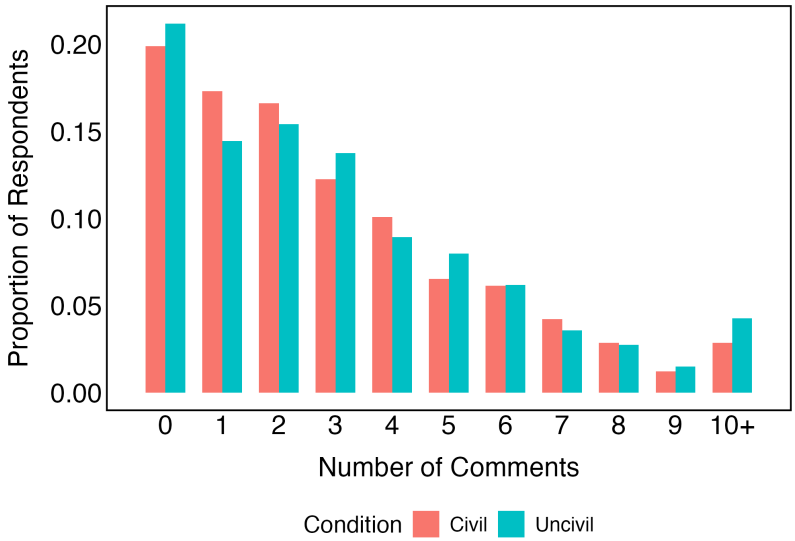



Figure S7: Distribution of post and comment counts by condition

Note: Proportions calculated within condition.

Table S7: Personal characteristics and online engagement by toxicity of posts and comments

| Variable | Post toxicity | | | | Comment toxicity | | | |
|---|---|---|---|---|---|---|---|---|
| | No posts | Low | Medium | High | No comments | Low | Medium | High |
| Age | 37.2 | 36.1 | 36.2 | 34.6 | 36.6 | 36.6 | 36.6 | 35.4 |
| College (%) | 59.0 | 60.7 | 58.0 | 50.0 | 60.0 | 62.3 | 56.7 | 52.9 |
| Female (%) | 68.6 | 67.2 | 65.9 | 59.6 | 72.3 | 72.4 | 65.4 | 56.4 |
| Race | | | | | | | | |
| ... Asian (%) | 4.0 | 5.5 | 2.8 | 4.5 | 3.7 | 4.2 | 5.0 | 2.1 |
| ... Black (%) | 13.1 | 14.9 | 12.2 | 11.5 | 14.3 | 18.3 | 11.8 | 8.3 |
| ... Hispanic (%) | 4.3 | 4.5 | 6.0 | 5.5 | 5.7 | 5.9 | 3.6 | 5.9 |
| ... Mixed (%) | 10.1 | 10.4 | 9.5 | 8.5 | 10.0 | 5.9 | 10.6 | 11.4 |
| ... White (%) | 68.5 | 64.7 | 69.5 | 70.0 | 66.3 | 65.7 | 69.0 | 72.3 |
| Partisanship | | | | | | | | |
| ... Democratic | 51.2 | 47.3 | 49.0 | 49.5 | 47.7 | 56.4 | 48.1 | 49.1 |
| ... Republican (%) | 39.0 | 42.8 | 41.5 | 39.5 | 40.0 | 37.0 | 40.3 | 43.6 |
| ... Independent (%) | 9.8 | 10.0 | 9.5 | 11.0 | 12.3 | 6.6 | 11.6 | 7.3 |
| Political interest | 3.3 | 3.3 | 3.3 | 3.3 | 3.1 | 3.3 | 3.3 | 3.4 |
| Social media use | 5.2 | 5.6 | 5.4 | 5.5 | 5.1 | 5.4 | 5.4 | 5.3 |
| Thermometers (0-100) | | | | | | | | |
| ... In-party | 67.0 | 67.8 | 65.9 | 66.5 | 67.3 | 66.4 | 67.6 | 65.0 |
| ... Out-party | 27.6 | 30.9 | 31.5 | 28.2 | 30.3 | 29.4 | 30.5 | 25.5 |
| Online engagement | | | | | | | | |
| ... # of posts | 0.0 | 1.1 | 1.2 | 1.3 | 0.4 | 0.6 | 0.7 | 0.9 |
| ... # of comments | 2.4 | 2.8 | 3.6 | 4.3 | 0.0 | 2.5 | 3.8 | 5.2 |
| ... Toxicity of posts | - | 0.0 | 0.0 | 0.2 | 0.1 | 0.0 | 0.1 | 0.1 |
| ... Toxicity of comments | 0.0 | 0.0 | 0.0 | 0.1 | - | 0.0 | 0.0 | 0.1 |
| ... Comfort sharing | 3.0 | 3.3 | 3.3 | 3.4 | 2.7 | 3.2 | 3.4 | 3.4 |

Note: Table shows personal characteristics and online engagement patterns of those with low, medium, or high post or comment toxicity, as well as those making no posts or comments
P-value is from t-test between the 'civil' and 'uncivil' conditions

# SI C Attrition Analysis

Figure S8 shows a flowchart of attrition across the stages of our study. Starting with the screening survey, we observed 1,593 of the 1,652 who started the survey make it to completion (96%), with 59 being forced to exit the survey due to failing to meet eligibility criteria (11), failing attention checks (39), or exiting voluntarily (9). We then observed 92 participants drop during the download phase as they did not make it to the beginning of the pre-treatment survey conducted within the app. The vast majority of those who started the pre-treatment survey (1,501) successfully completed it (1,491; $>$ 99%), with only 10 respondents dropping at this stage. Next, we see some attrition at the newsfeed stage as 24 of the 1,491 who finished the pre-treatment survey (1.6%) did not reach the beginning of the post-treatment survey. Attrition through the last stage of the study was low with only five of the 1,467 who started the post-treatment survey (0.3%) failing to make it to the end. In total, we were left with 1,462 completes, with 734 in the 'civil' condition and 728 in the 'uncivil' condition. We remove one respondent for data quality purposes as pre-registered,

leaving a final analytic sample of 1,461.

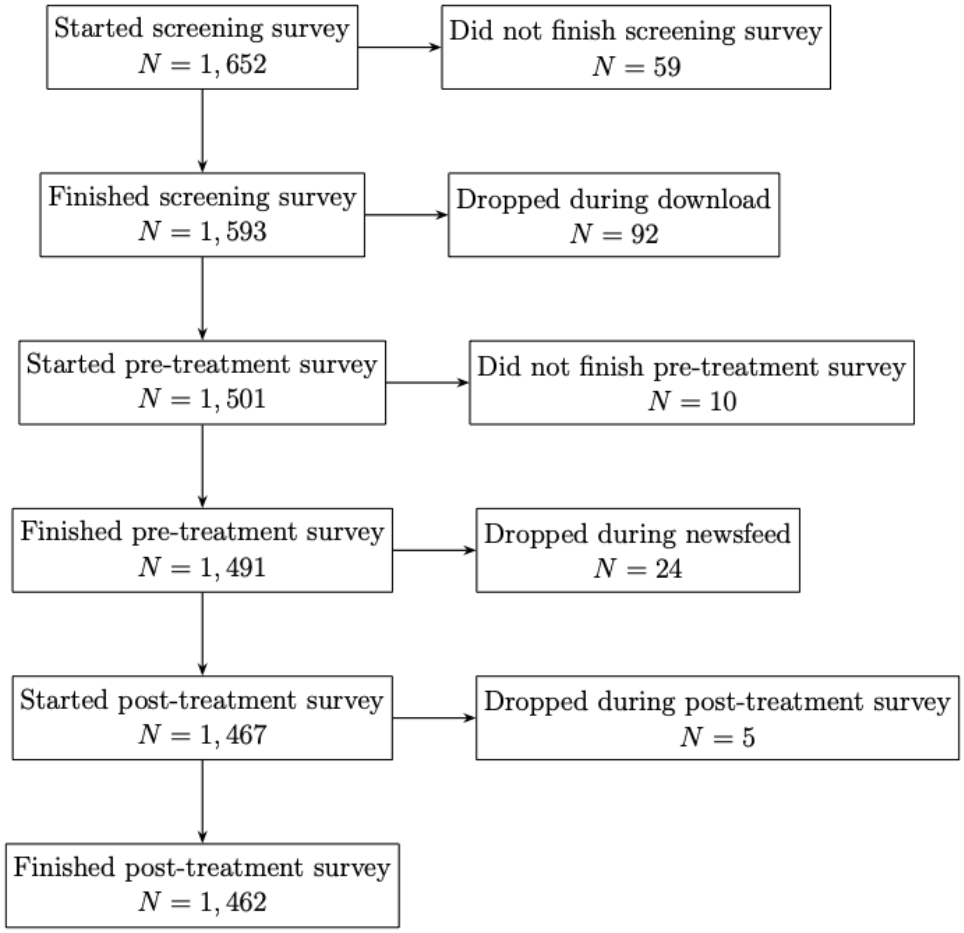



Figure S8: Flowchart of attrition through full study

# SI D Supplemental Tables, Figures, and Analyses

Table S8: ATE of assignment to uncivil newsfeed (H1-H3)

| | H1: Comfort sharing | H2a: # of posts | H2b: # of comments | H2c: # of out-party comments | H3a: Toxicity of posts | H3b: Toxicity of comments | H3c: Toxicity of out-party comments |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Uncivil newsfeed | −0.293** | 0.072* | 0.188 | −0.095 | 0.371** | 0.356** | 0.410** |
| | (0.052) | (0.037) | (0.154) | (0.085) | (0.070) | (0.058) | (0.077) |
| (Constant) | 0.146** | 0.624** | 2.946** | 1.133** | −0.194** | −0.176** | −0.192** |
| | (0.037) | (0.026) | (0.109) | (0.060) | (0.050) | (0.041) | (0.053) |
| Observations | 1,461 | 1,461 | 1,461 | 1,317 | 801 | 1,156 | 651 |
| Adjusted $R^2$ | 0.021 | 0.002 | 0.0003 | 0.0002 | 0.033 | 0.031 | 0.040 |

Note: Table shows coefficient estimates of assignment to uncivil newsfeed on outcomes associated with Hypotheses 1-3
†p<0.1; *p<0.05; **p<0.01; one-tailed
Standard errors in parentheses
Toxicity in Columns 5-7 operationalized as the mean toxicity of a respondent's posts or comments, respectively. These analyses are necessarily limited to respondents providing at least one post, comment, or out-party comment, respectively
Leaners included as partisans when operationalizing the toxicity of out-party comments in Column 7.

Table S9: ATE of assignment to uncivil newsfeed (H4-H7)

| | H4a: In-party thermometer | H4b: Out-party thermometer | H5: Perceived polarization | H6: Trust in government | H7: Democratic satisfaction |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Uncivil newsfeed | 0.020 | −0.085** | 0.007 | 0.015 | −0.026 |
| | (0.029) | (0.025) | (0.037) | (0.025) | (0.028) |
| Pretreatment: In-party therm. | 0.043** | | | | |
| | (0.001) | | | | |
| Pretreatment: Out-party therm. | | 0.043** | | | |
| | | (0.001) | | | |
| Pretreatment: Perceived polarization | | | 0.354** | | |
| | | | (0.009) | | |
| Pretreatment: Trust in gov. | | | | 1.102** | |
| | | | | (0.016) | |
| Pretreatment: Democratic satisfaction | | | | | 0.578** |
| | | | | | (0.010) |
| (Constant) | −2.857** | −1.210** | −1.222** | −2.382** | −1.611** |
| | (0.052) | (0.025) | (0.040) | (0.038) | (0.034) |
| Observations | 1,317 | 1,317 | 1,457 | 1,460 | 1,461 |
| Adjusted $R^2$ | 0.725 | 0.797 | 0.509 | 0.770 | 0.707 |

Note: Table shows coefficient estimates of assignment to uncivil newsfeed on outcomes associated with Hypotheses 4-7
†p<0.1; *p<0.05; **p<0.01; one-tailed
Standard errors in parentheses
Leaners included as partisans when operationalizing the in-party and out-party thermometer ratings in Columns 1 and 2.

Table S10: ATE of assignment to uncivil newsfeed on within-subject change outcomes (H4-H7)

| | H4a: ΔIn-party thermometer | H4b: ΔOut-party thermometer | H5: ΔPerceived polarization | H6: ΔTrust in government | H7: ΔDemocratic satisfaction |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Uncivil newsfeed | 0.018 | −0.173** | −0.018 | 0.014 | −0.032 |
| | (0.055) | (0.055) | (0.052) | (0.052) | (0.052) |
| (Constant) | −0.009 | 0.086* | 0.009 | −0.007 | 0.016 |
| | (0.039) | (0.039) | (0.037) | (0.037) | (0.037) |
| Observations | 1,317 | 1,317 | 1,457 | 1,460 | 1,461 |
| Adjusted $R^2$ | -0.001 | 0.007 | -0.001 | -0.001 | -0.0004 |

Note: Table shows coefficient estimates of assignment to uncivil newsfeed on within-subject change outcomes associated with Hypotheses 4-7
†p<0.1; *p<0.05; **p<0.01; one-tailed
Standard errors in parentheses
Leaners included as partisans when operationalizing the in-party and out-party thermometer ratings in Columns 1 and 2.

Table S11: ATE of assignment to uncivil newsfeed on post and comment toxicity with random intercepts

| | Post toxicity | Comment toxicity | Out-party comment toxicity |
|---|---|---|---|
| | (1) | (2) | (3) |
| Uncivil condition | 0.309** | 0.232** | 0.322** |
| | (0.064) | (0.037) | (0.058) |
| (Constant) | −0.163** | −0.151** | −0.204** |
| | (0.047) | (0.027) | (0.040) |
| Random intercepts? | *Yes* | *Yes* | *Yes* |
| Observations | 951 | 4,356 | 1,412 |
| Akaike Inf. Crit. | 2,691.657 | 12,071.040 | 3,786.698 |

Note: Table shows coefficient estimates of assignment to uncivil newsfeed on toxicity of posts and comments, estimated with random intercepts
†p<0.1; *p<0.05; **p<0.01; one-tailed
Standard errors in parentheses
Unit of analysis is the respondent-post (Column 1) and respondent-comment (Columns 2-3)
Toxicity measures have been standardized

Table S12: ATE of assignment to uncivil newsfeed on conversational features of posts

| | Politeness (reversed) | VADER sentiment (reversed) | Identity attack | Insult | Profanity |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Uncivil newsfeed | 0.121$^{\dagger}$ | 0.037 | 0.227$^{**}$ | 0.341$^{**}$ | 0.299$^{**}$ |
| | (0.070) | (0.071) | (0.070) | (0.070) | (0.070) |
| (Constant) | −0.063 | −0.020 | −0.119$^{*}$ | −0.179$^{**}$ | −0.157$^{**}$ |
| | (0.051) | (0.051) | (0.051) | (0.051) | (0.051) |
| Observations | 805 | 805 | 801 | 801 | 801 |
| Adjusted $R^2$ | 0.002 | -0.001 | 0.012 | 0.028 | 0.021 |

Note: Table shows coefficient estimates of assignment to uncivil newsfeed on conversational features of posts
$^{\dagger}$p<0.1; $^{*}$p<0.05; $^{**}$p<0.01; two-tailed
Standard errors in parentheses
Unit of analysis is the respondent

Table S13: ATE of assignment to uncivil newsfeed on conversational features of comments

| | Politeness (reversed) | VADER sentiment (reversed) | Identity attack | Insult | Profanity |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Uncivil newsfeed | 0.126$^{*}$ | 0.116$^{*}$ | 0.229$^{**}$ | 0.356$^{**}$ | 0.269$^{**}$ |
| | (0.059) | (0.059) | (0.058) | (0.058) | (0.058) |
| (Constant) | -0.062 | -0.057 | -0.113$^{**}$ | -0.176$^{**}$ | -0.133$^{**}$ |
| | (0.041) | (0.041) | (0.041) | (0.041) | (0.041) |
| Observations | 1,157 | 1,157 | 1,156 | 1,156 | 1,156 |
| Adjusted $R^2$ | 0.003 | 0.002 | 0.012 | 0.031 | 0.017 |

Note: Table shows coefficient estimates of assignment to uncivil newsfeed on conversational features of comments
$^{\dagger}$p<0.1; $^{*}$p<0.05; $^{**}$p<0.01; two-tailed
Standard errors in parentheses
Unit of analysis is the respondent

Table S14: ATE of assignment to uncivil newsfeed on conversational features of posts with random intercepts

| | Politeness (reversed) | VADER sentiment (reversed) | Identity attack | Insult | Profanity |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Uncivil newsfeed | 0.108$^{\dagger}$ | 0.023 | 0.223$^{**}$ | 0.296$^{**}$ | 0.255$^{**}$ |
| | (0.065) | (0.065) | (0.071) | (0.064) | (0.064) |
| (Constant) | −0.057 | −0.012 | −0.113$^{*}$ | −0.155$^{**}$ | −0.134$^{**}$ |
| | (0.047) | (0.047) | (0.051) | (0.047) | (0.047) |
| | (0.224) | (0.779) | (0.020) | (0.001) | (0.003) |
| *Random Intercepts?* | Yes | Yes | Yes | Yes | Yes |
| Observations | 962 | 962 | 951 | 951 | 951 |
| Akaike Inf. Crit. | 2,742.897 | 2,745.586 | 2,694.866 | 2,693.538 | 2,698.996 |

Note: Table shows coefficient estimates of assignment to uncivil newsfeed on conversational features of posts, estimated with random intercepts
$^{\dagger}$p<0.1; $^{*}$p<0.05; $^{**}$p<0.01; two-tailed
Standard errors in parentheses
Unit of analysis is the respondent-post

Table S15: ATE of assignment to uncivil newsfeed on conversational features of comments with random intercepts

| | Politeness (reversed) | VADER sentiment (reversed) | Identity attack | Insult | Profanity |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Uncivil newsfeed | 0.088$^{**}$ | 0.096$^{**}$ | 0.117$^{**}$ | 0.227$^{**}$ | 0.164$^{**}$ |
| | (0.033) | (0.034) | (0.032) | (0.037) | (0.036) |
| (Constant) | −0.045$^{\dagger}$ | −0.055$^{*}$ | −0.065$^{**}$ | −0.139$^{**}$ | −0.104$^{**}$ |
| | (0.023) | (0.025) | (0.023) | (0.026) | (0.025) |
| *Random Intercepts?* | Yes | Yes | Yes | Yes | Yes |
| Observations | 4,396 | 4,396 | 4,356 | 4,356 | 4,356 |
| Akaike Inf. Crit. | 12,467.890 | 12,440.320 | 12,351.040 | 12,109.940 | 12,205.700 |

Note: Table shows coefficient estimates of assignment to uncivil newsfeed on conversational features of comments, estimated with random intercepts
$^{\dagger}$p<0.1; $^{*}$p<0.05; $^{**}$p<0.01; two-tailed
Standard errors in parentheses
Unit of analysis is the respondent-comment

Table S16: Comparing features of user-generated content by condition

| Content Type | Measure | Uncivil | Civil | Diff. | P-value |
|---|---|---|---|---|---|
| Comments | Toxicity | 0.076 | 0.047 | 0.029 | 0.000 |
| Comments | Identity attack | 0.009 | 0.006 | 0.003 | 0.000 |
| Comments | Insult | 0.046 | 0.023 | 0.023 | 0.000 |
| Comments | Profanity | 0.041 | 0.026 | 0.014 | 0.000 |
| Comments | Politeness† | 0.009 | −0.010 | 0.020 | 0.004 |
| Comments | VADER† | −0.149 | −0.189 | 0.041 | 0.001 |
| Posts | Toxicity | 0.085 | 0.051 | 0.035 | 0.000 |
| Posts | Identity attack | 0.014 | 0.008 | 0.006 | 0.004 |
| Posts | Insult | 0.052 | 0.024 | 0.028 | 0.000 |
| Posts | Profanity | 0.044 | 0.024 | 0.020 | 0.000 |
| Posts | Politeness† | 0.011 | −0.012 | 0.024 | 0.093 |
| Posts | VADER† | −0.144 | −0.154 | 0.010 | 0.724 |

Note: Table shows mean values of various conversational features of posts and comments across uncivil and civil conditions, along with their difference and associated $p$-value
$p$-values from two-tailed tests given exploratory nature of analyses
† Politeness and VADER sentiment are reversed-coded so that higher values = more impolite or more negative sentiment.

Table S17: ATE of assignment to uncivil newsfeed on commenting behavior

| | Commenting: | | | Toxicity: | | |
|---|---|---|---|---|---|---|
| | In-Party | Out-Party | Non-Political | In-Party | Out-Party | Non-Political |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Uncivil newsfeed | −0.063* | −0.062* | 0.037 | 0.030** | 0.037** | 0.003 |
| | (0.028) | (0.028) | (0.023) | (0.008) | (0.007) | (0.006) |
| (Constant) | 0.518** | 0.529** | 0.255** | 0.045** | 0.038** | 0.043** |
| | (0.019) | (0.019) | (0.016) | (0.005) | (0.005) | (0.004) |
| *Random Intercepts?* | No | No | No | Yes | Yes | Yes |
| Observations | 1,317 | 1,317 | 1,461 | 1,219 | 1,412 | 559 |
| Adjusted $R^2$ | 0.003 | 0.003 | 0.001 | | | |
| Akaike Inf. Crit. | | | | -1,697.549 | -2,282.086 | -1,398.812 |

Note: Table shows coefficient estimates of assignment to uncivil newsfeed on comments on posts and toxicity of comments, by post type. Models in Columns 4-6 estimated with random intercepts
†p<0.1; *p<0.05; **p<0.01; two-tailed
Standard errors in parentheses
Unit of analysis is the respondent in Columns 1-3 and the respondent-comment in Columns 4-6.

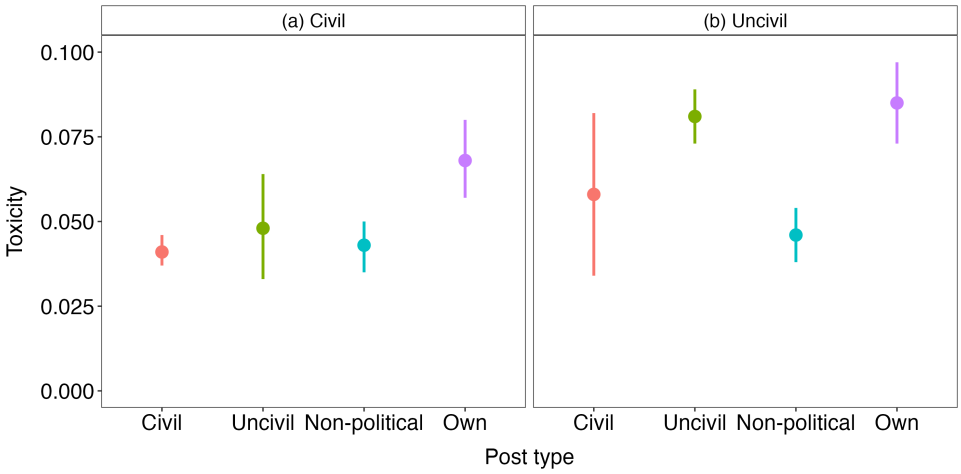



Figure S9: Comment toxicity by post type and condition

Note: Points and error-bars represent estimated and 95% confidence intervals, respectively.

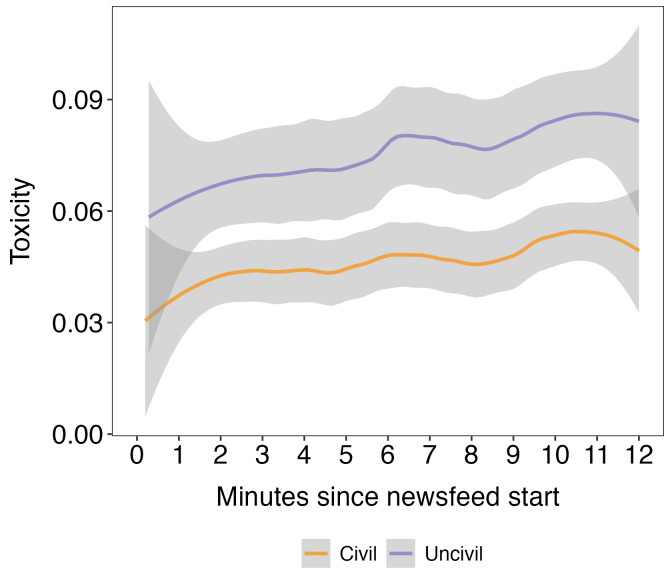



Figure S10: Comment toxicity across the 12-minute newsfeed session by condition

Note: Grey bands around the toxicity estimates use LOESS smoothing method.