Promoting Online Civility Through Platform Architecture

Jisu Kim, Curtis McDonald, Paul Meosky, Matt Katsaros, and Tom Tyler

Abstract. This study tests whether the architecture of a social media platform can encourage conversations among users to be more civil. It was conducted in collaboration with Nextdoor, a networking platform for neighbors within a defined geographic area. The study involved: (1) prompting users to move popular posts from the neighborhood-wide feed to new groups dedicated to the topic and (2) an experiment that randomized the announcement of community guidelines to members who join those newly formed groups. We examined the impact of each intervention on the level of civility, moral values reflected in user comments, and user's submitted reports of inappropriate content. In a large quantitative analysis of comments posted to Nextdoor, the results indicate that platform architecture can shape the civility of conversations. Comments within groups were more civil and less frequently reported to Nextdoor moderators than the comments on the neighborhood-wide posts. In addition, comments in groups where new members were shown guidelines were less likely to be reported to moderators and were expressed in a more morally virtuous tone than comments in groups where new members were not presented with guidelines. This research demonstrates the importance of considering the design, structure, and affordance of the online environment when online platforms seek to promote civility and other pro-social behaviors.

1 Introduction

Social media platforms have become increasingly central to the social, civic, political, and economic issues affecting communities around the world (Jin 2015). Such platforms can provide a new mechanism for restoring social connections and enhancing traditional communities (Phua, Jin, and Kim 2017). For example, Facebook, now Meta, touts its mission as "giv[ing] people the power to build community and bring[ing] the world closer together" (Meta Platforms, Inc. 2022). Twitter's stated purpose is to "serve the public conversation" by providing "a free and safe space to talk" (Twitter, Inc. 2022). And Youtube's mission is "to give everyone a voice and show them the world" (YouTube 2022).

doi:10.54501/jots.v1i4.54

What happens on platforms is not limited to the digital sphere. Social media conversations also affect the offline communities in which platform users live, work, and play. Online platforms such as Facebook can provide a way for people to seek emotional support when faced with a cancer diagnosis (Bender, Jimenez-Marroquin, and Jadad 2011), while sites such as Tinder can facilitate lifelong romantic partnerships. At the local level, platforms such as Front Porch Forum can build online communities to allow neighbors to interact with each other to discuss anything from improving trash collection to spending local tax dollars. Similarly, Nextdoor is an online platform dedicated to "bringing neighbors and organizations together, [to] cultivate a kinder world where everyone has a neighborhood they can rely on" (Nextdoor 2022a). Thus, whether connecting users to friends across the world or neighbors down the street, these platforms facilitate diverse forms of collaborative and productive social interactions.

There is no shortage of high-profile examples demonstrating the unintended consequences and negative externalities that can occur when the entire world connects online. Online bullying, cyber stalking, expressions of hate speech, and coordinated disinformation campaigns can have negative psychological and behavioral implications for users (Gahagan, Vaterlaus, and Frost 2016; Rieger et al. 2021). For example, community members in the groups we studied on Nextdoor faced racist and homophobic slurs, belittling scorn, and even overt threats. Recognition of this possibility has driven significant investments and efforts by researchers and online platforms to identify and regulate various forms of undesirable speech and behaviors. Trying to moderate undesirable content to reduce harm is an ongoing challenge.

When social media platforms emerged, many were initially viewed as providing an opportunity to connect with people around the world. The goal was not harm reduction, but to build social relations both on and off platforms, thereby improving individual psychological well-being and ameliorating the social, political, and economic vitality of real-world communities. However, scholars have focused on examining platform design and regulation that seeks to reduce negative psychological and behavioral impacts for users (Jhaver et al. 2019; Tyler et al. 2021); comparatively little research on encouraging positive behaviors on social media platforms. Hence, the goal of this study is to refocus on the potential of social media platforms to promote individual and community well-being. This research was conducted in collaboration with the neighborhood-based social media platform Nextdoor. It analyzes comments and platform behavior to test whether platform architecture—the designs, affordances, and structures that shape user interactions on a platform—can positively influence the civility and moral values of online discussions.

Specifically, this study examines the relationship of two interventions on the neighborhood-based social media platform Nextdoor—(1) creating a group dedicated to discussing a given topic and (2) announcing (or not) community guidelines to new members of the group—and the civility and moral values of online discussions. For the first intervention, authors of posts that generated active conversations within the community were asked if they would like to create a group dedicated to the topic of their post. For those who chose to create a new group, we conducted a pretest-posttest analysis to examine what happens when these popular conversations move into a group setting by comparing the levels of civility and moral values exhibited in the user comments on the original post (Neighborhood Post Comments) with the comments made in the newly formed group (Group Posts Comments) on Nextdoor (see Table 1 on page 6 for definitions). In the second intervention, we used a randomized controlled experiment to test the effect of showing basic community guidelines to users joining these newly formed groups on the civility of group interactions.

The results demonstrate that platform architecture can be used to encourage users to

engage in more civil interactions. The comments made in the new groups were associated with higher levels of civility, more virtuous moral values, and fewer incidents of user reports relative to the comments on the original Neighborhood Post. In addition, showing guidelines to new group members caused an increase in comments with more virtuous moral values and a decrease in user reports of comments. These findings demonstrate the need for social media platforms to design and architect platforms that can clearly communicate expectations and norms to users to encourage more civil interactions.

2 Literature Review

2.1 The Effect of Social Media Platform Architecture and Affordances on User Discussions

Thus far, scholars have explored how various aspects of social media platform architecture—the design components, affordances, and structures—influence and shape users' behaviors in online and social media environments. First, a group of scholars have focused on the existence of anonymity on online platforms affects the civility of users' discussions (Coe, Kenski, and Rains 2014; Santana 2013; Ruiz et al. 2011). Coe, Kenski, and Rains (2014) and Ruiz et al. (2011) found that user registration (which connects individual usernames with personally identifiable information) discouraged hostile comments on newspaper websites. Santana (2013) also found that non-anonymous comments were more civil than anonymous comments in online newspaper discussion forums. Rowe (2014) compared comments on political news between the *Washington Post* site and the *Washington Post* Facebook page and found that comments on Facebook were more civil and polite than on the news site due to the lack of anonymity. However, Hille and Bakker (2014) found that anonymous comments on news sites were more elaborate than non-anonymous comments on Facebook. These mixed results suggest that anonymity itself is ethically neutral (Reader 2012) and should not be the only factor affecting users' civil interactions.

In addition to anonymity, scholars have focused on how other affordances can enable or constrain user behaviors and affect civil discussions (Jaidka, Zhou, and Lelkes 2019). Jaidka, Zhou, and Lelkes (2019) analyzed a large volume of tweet responses to U.S. politicians and found that doubling the character limit of tweets encouraged users to engage in less uncivil, more polite, and constructive discussions. Seering et al. (2019) conducted a survey experiment and found that CAPTCHAs containing stimuli designed to prime positive emotions and mindsets could increase the positivity of sentiments and the levels of complexity and social connectedness in participants' comments on politically charged comment threads.

Other studies have investigated how online platforms' specific moderation policies affect users' civil discussions (Ksiazek 2015; Jhaver et al. 2019; Lampe et al. 2014; Matias 2019; Ribeiro, Cheng, and West 2022; Tyler et al. 2021). For example, Ksiazek (2015) analyzed online public discussions on U.S. news organization websites and found that pre-moderation (i.e., an automatic filter system) and post-moderation (i.e., flagging) increased civility, while offering a private messaging option boosted hostility. Lampe et al. (2014) tested how distributed moderation systems affect civil conversations on Slashdot, a membership-based online news and discussion site. Ribeiro, Cheng, and West (2022) assessed community participation in Facebook groups after the group admins turned on post approvals (requiring a group admin to review and approve posts before they were shown to group members); they found that adopting the post approvals feature reduced the number of posts to the group, however these posts received more

comments and were reported less often by community members.

Finally, Matias (2019) conducted a large-scale field experiment to examine how community rules influence who chooses to join a group and how they behave on Reddit (r/science). Their results demonstrated that announcing the community guidelines increased norm compliance among first-time participants in the group. Jhaver et al. (2019) also found that users who had posts removed from Reddit were less likely to have future posts removed when provided with explanations. Similarly, Tyler et al. (2021) determined that users who were provided education about platform rules in the week following their post removal were less likely to have future posts removed. Katsaros, Yang, and Fratamico (2022) conducted a field experiment on Twitter that asked users who posted content with offensive language if they would like to reconsider their post. This intervention resulted in 31% of participants editing or deleting their post.

While much prior research has suggested that platform architecture and affordances would affect online user discussions across different platforms, our study explores whether similar approaches can be used to shape civil and moral interactions on the neighborhood-based platform Nextdoor.

2.2 Civility in Online Discussions

Previous studies have employed different approaches to explore the abstract concept of civility as a norm, custom, moral obligation, strategy, formality, set of requirements, and mechanism (Calhoun 2000; Papacharissi 2004; Waldron 2013). For instance, Whitman (2000) suggested that civility is related to showing respect to others, which requires individuals to acknowledge each other as equals. Waldron (2013) defined civility as the hard work of staying present in the discussion, even when facing deep-rooted disagreement. Civility has also been linked to politeness (Walter and Lipsitz 2021), but unlike mere politeness, civility entails explicitly affirming another's values or ideas, even those one finds disagreeable (Han, Brazeal, and Pennington 2018). There is a general consensus that civility is a necessary component to maintaining and promoting an effective democracy by respecting different views (Smith and Bressler 2013).

However, establishing a definition of civility is challenging because the notion involves conforming to socially created norms or rules (Calhoun 2000). As Papacharissi (2004) suggested, civility can be regarded as respect for the collective traditions of democracy that are accepted by particular local cultures. Because the concept is tied to particular, contingent, and contextually specific social rules regarding behavior, critiques have pointed out how civility as a concept reinforces the status quo and imposes the norms of a dominant group on minorities, which can be morally ambiguous Jamieson et al. (2017) and Zurn (2013).[1]

Still, civility does not only involve following a particular norm or rule and conforming to the local culture. It should also be regarded as a general moral virtue itself in communicating with others, such as demonstrating tolerance and respect toward others beyond conforming to a specific set of social rules (Calhoun 2000). Civility can be tied to a specific culture by avoiding unnecessarily disrespectful words toward the discussion participants. Many contemporary efforts to conceptualize civility focus on the nature of deliberation and discussions involving appropriate forms of disagreement on moral matters (Jamieson et al. 2017). A willingness to listen to others based on tolerance and respect of others in discussions is also an important moral aspect of civility (Rawls 1996).

1. For this reason, we will discuss the moral ambiguity of civility in greater depth in a future article on the methodological lessons learned from developing our civility codebook, currently titled "Civil to whom?: Measuring online civility through iterative coding."

By focusing more on the mode of interaction, our study defines civility as demonstrating tolerance and respect toward others in the discussion—even in the face of those who have opposing views and ideas. The extant literature on civility has highlighted these two prominent and general components of civility—tolerance and respect toward others (Han, Brazeal, and Pennington 2018; Waldron 2013; Whitman 2000). Specifically, tolerance involves recognizing that others have different views and ideas and providing a neutral environment in which citizens can exchange viewpoints. By acknowledging that others can have different viewpoints, tolerance enables us to disagree on controversial subjects and debate these differences in a civil and nonviolent manner. Such debates are necessary for resolving disputed issues and developing fair and enlightened public policies. Respect toward others is defined as acknowledging the autonomy and dignity of every citizen as a free and equal human being, regardless of his or her specific traits or opinions. In other words, to uphold the value of respect toward others, online and offline discussion communities should be inclusive, and every member should be able to share his or her ideas and challenge those of others (Benn and Benn 1988).

As use of social media has grown, particularly platforms that allow anonymous usage, so too has the prevalence of rude, uncivil discussion and hate speech (Santana 2013). Prior studies in this area have analyzed the degree of civil and uncivil language conveyed in online discussions (Coe, Kenski, and Rains 2014; Papacharissi 2004), various factors affecting individuals' uncivil discussions online (Blom et al. 2014), and the effect of online platform policies, including user registration, anonymity, and moderation of civil discussion in online spaces (Ksiazek 2015; Lampe et al. 2014; Santana 2013). Although there has been considerable research on the nature and extent of incivility online, there is startlingly little scholarship on the prevalence and mechanisms of civility online. More importantly, scholars have commonly operationalized civility as the absence of incivility (Papacharissi 2004; Santana 2013) or focused on incivility or aggressive language use in online discussions (Coe, Kenski, and Rains 2014; Ksiazek 2015) rather than focusing on the original concept itself. This study takes a more pro-social approach by (1) developing our own codebook to measure civility and (2) defining and measuring civility as a moral virtue.

2.3 Moral Values in Online Discussions

Based on cultural psychology, the Moral Foundation Theory (MFT) defines moral values (i.e., morality) as a set of values, practices, institutions, and psychological mechanisms that suppress selfishness and regulate social life for group cohesiveness and harmony (Haidt 2008). The core of MFT is that different cultures share basic moral values. Haidt and Joseph (2004) originally identified four "moral modules" that they later refined into five "moral foundations": (1) care/harm, (2) fairness/cheating, (3) loyalty/betrayal, (4) authority/subversion, and (5) sanctity/degradation (Haidt and Graham 2007). Graham and Haidt (2012) has since added a sixth foundation, liberty/oppression, and others have recommended additional foundations such as equality (as distinct from proportionality).

The care/harm foundation is related to basic concerns about others' suffering by caring, nurturing, and protecting vulnerable individuals (Graham, Haidt, and Nosek 2009; Haidt, Graham, and Joseph 2009). The fairness/cheating foundation is based on concerns about meritocracy—and, to a lesser extent, equality—and generates the idea of justice (Haidt, Graham, and Joseph 2009; Graham and Haidt 2012). The loyalty/betrayal foundation is closely connected to commitment to and self-sacrifice for the sake of a group. The authority/subversion foundation is linked to the social order and obligations of hierarchical relationships, including deference and respect for tradition, leaders, and hierarchical organization (Haidt, Graham, and Joseph 2009). The sanctity/degrada-

tion foundation is based on concerns about "physical and spiritual contagion, including virtues of chastity, wholesomeness, and control of desires" (Haidt, Graham, and Joseph 2009). The first three foundations are closely connected to individuals' freedom and rights (i.e., individualizing foundations), while the other three bind individuals to a group or collective (i.e., binding foundations). Walter and Lipsitz (2021) suggests that those who hold individualizing foundations tend to have a stronger emotional response to uncivil discussion than those who hold binding foundations. The new liberty/oppression foundation is based on the resentment one feels towards domination, bullying, and oppression (Graham and Haidt 2012).

Previous linguistic and computer science studies have analyzed large volumes of textual data to develop moral dictionaries (Araque, Gatti, and Kalimeri 2020; Hoover et al. 2021) and examined the types of moral values represented in social media users' discussions (Grover et al. 2019). In addition, scholars have investigated how individual characteristics (especially political predisposition) affect the endorsement of each moral value (Graham, Haidt, and Nosek 2009; Haidt and Graham 2007). More importantly, moral values have been regarded as the set of values and practices that suppress selfishness and regulate civil life for group cohesiveness and harmony (Haidt 2008), and eventually lead people to engage in pro-social behaviors (Nilsson, Erlandsson, and Västfjäll 2016; Welsch 2020). Thus, we use MFT to measure the degree of moral values in comments as an indicator of social media users' pro-social behavior on the platform.

Table 1: Definitions of key terms used throughout this paper.

Term in Paper	Definition
Neighborhood Post	General posts made by Nextdoor users. These posts can be viewed and commented on by neighbors of the post author; these posts appear on neighbors' main feed when they log on to Nextdoor.
Neighborhood Post Comment	Comments made on a Neighborhood Post. Because Neighborhood Posts are only visible to neighbors of the Neighborhood Post author, the author of the Neighborhood Post Comment must be a neighbor of the Neighborhood Post author.
Group Post	Posts made within a group on Nextdoor. In this study, we asked the authors of popular Neighborhood Posts if they would like to create a group to continue conversation on that topic. While Neighborhood Posts are open for anyone in the neighborhood to view and comment on, Groups on Nextdoor can be a way to discuss a specific topic among a smaller subset of the neighborhood. These groups can be open (anyone in the neighborhood can join the group and participate in discussions) or private (anyone in the neighborhood can view the group and request to join, but the group admin must approve a membership request before a member can participate in discussions)(Nextdoor 2022b)
Group Comment	Comments made on a Group Post

3 Hypotheses

Building on previous literature, we analyzed comments and platform behavior to test whether an online platform's design architecture can positively influence the degree to which discussions among users are civil and reflect moral values. We did so through one pretest-postest analysis and one field experiment conducted in collaboration with the neighborhood-based social media platform Nextdoor involving two interventions.

First, we analyzed the result of Nextdoor encouraging authors of highly commented Neighborhood Posts to create a new group dedicated to the topic of their post. For this intervention, we used a single-group pretest-posttest analysis design to compare comments made on the original Neighborhood Post to comments made in the newly formed Group to test the following hypotheses:

- *H1a:* Group Post Comments are more civil than comments on the corresponding Neighborhood Posts.
- *H1b:* Group Post Comments include more virtuous moral values than comments on the corresponding Neighborhood Posts.
- *H1c:* Group Post Comments have fewer comments reported by users than comments on the corresponding Neighborhood Posts.

In the second intervention we used a random assignment experiment design to test the effect of announcing community guidelines to new group members on interactions within that newly formed group. Previous findings indicate that announcing guidelines or rules can affect users' behaviors (Matias 2019; Jhaver et al. 2019; Tyler et al. 2021). While many prior studies focus on reinforcing a particular set of rules aimed at reducing antisocial behaviors, we investigate whether similar approaches can be used to shape civil interactions and increase pro-social behaviors. This part of the experiment tests our second set of hypotheses:

- *H2a:* Providing members with guidelines before they enter a newly formed group will result in more civil Group Comments in groups with guidelines compared to Group Comments in groups without guidelines.
- *H2b:* Providing members with guidelines before entering a newly formed group will result in more virtuous moral values in Group Comments in groups with guidelines compared to Group Comments in groups without guidelines.
- *H2c:* Providing members with guidelines before entering a newly formed group will result in fewer user reports of Group Comments in groups with guidelines compared to Group Comments in groups without guidelines.

4 Methods

4.1 Study Design

In our study conducted in collaboration with Nextdoor, we tested the influence of two important architectural features: (1) prompting authors of highly engaging Neighborhood Posts to create a new group dedicated to a specific issue and (2) announcing community guidelines to new members of this newly formed group. For clarity, we have provided a table which defines terms for specific content types analyzed in this study shown in Table 1 on the preceding page.

Nextdoor's privacy boundaries are designed to replicate physical geographic neighbor-

hood boundaries: users on Nextdoor can only see the posts, comments, and activities from their actual neighbors. To register for a Nextdoor account, users must confirm their location through a physical piece of mail sent to their address. As a result, interactions on this platform can connect people who share membership in a particular geographical area. The platform's goal is to leverage this shared membership to create positive and constructive interactions about shared problems and issues in users' communities.

Neighborhood Posts can only be seen and commented on by other users in the post author's neighborhood. While anyone in the neighborhood can engage with Neighborhood Posts, groups can be a way to discuss a specific topic among a smaller subset of the neighborhood. These groups can be open (anyone in the neighborhood can join the group and participate in discussions) or private (anyone in the neighborhood can view the group and request to join, but the group admin must approve a membership request before a member can participate in discussions) (Nextdoor 2022b).

4.2 Intervention 1: New Group Formation

For our group formation intervention, when any Neighborhood Post received its 70th Neighborhood Post Comment within our study, the platform messaged the Neighborhood Post author. Neighborhood Posts with 70 Neighborhood Post Comments indicated that the conversation was of significant interest to the neighborhood.[2] This message indicated that their post appeared to be generating a lot of conversation within the community, and invited the post author to create a new group dedicated to the issue(s) discussed in the post. For this intervention, we used a single-group pretest-posttest analysis design to compare the first 70 Neighborhood Post Comments (before the Neighborhood Post author was asked to create a new group) to Group Comments made in the newly formed group.

4.3 Intervention 2: Guidelines vs. No Guidelines

For the second intervention, all of these newly formed groups were randomly assigned into one of two conditions: Guidelines or No Guidelines. In the Guidelines condition, any new member joining the newly formed group was shown a set of guidelines; those in the No Guidelines condition were not shown any guidelines. These guidelines were minimally intrusive on a new member's experience (a single page shown before entering the group for the first time). Members were provided four short guidelines designed to promote more civil interactions within the group (see Figure 1 on the next page). The guidelines reflect the four antecedents of procedural justice: voice, respect, neutrality, and trustworthiness (Tyler, Jackson, and Bradford 2014).

4.4 Measures

4.4.1 Civility

To measure the level of civility in user discussions on Nextdoor, we developed a codebook through a literature review and an iterative labeling process (more details available in the supplementary material). A team of 14 undergraduate students used the codebook to label 7,816 comments over a 2-month period. The final codebook consisted of 13 civil labels and 13 uncivil labels. As noted above, in building our codebook we did not define civility as merely the absence of uncivil language (and visa versa). Two binary

2. The range of topics discussed in these posts were broad, examples include: social discussions (getting to know you, organizing parties or meet-ups, etc); Donations/Charity/Requests for Help or Prayers; local news; Politics; Race; Policing; Pets; and Schools/Education.

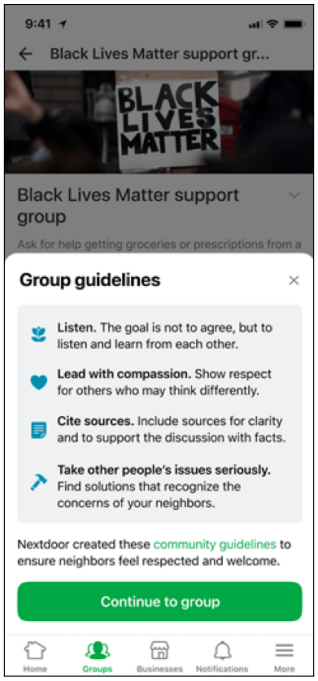

Figure 1: Group Guidelines shown to members entering newly formed groups on Nextdoor

classifications were generated for each comment reviewed: (1) a civil classification (either "civil" or "not civil") indicating the presence (or lack thereof) of civil discussion and (2) an uncivil classification (either "uncivil" or "not uncivil") indicating the presence (or lack thereof) of uncivil discussion. As such, a single comment can contain any two combinations of these four labels. Figure 2 on the next page displays a confusion matrix showing this overlap for all labeled comments. It shows that nearly two-thirds of all comments we labeled were either only civil (42.0%) or only uncivil (17.4%), while over one-third contained neither (37.3%); few comments contained both civil and uncivil language (3.3%).

Three student labelers blindly reviewed each comment, and a majority vote of the three was used to classify the comment. In total, 7,816 individual comments were labeled, totaling 23,448 distinct comment—labeler pairs. S1 in the supplementary materials describes the coding procedures in more detail and provides inter-rater reliability measures.

4.4.2 Moral Values

Labeling individual comments using our civility guidelines can provide very high fidelity data, but it comes at a high operational cost and results in only a small amount of data being analyzed, which limits our ability to detect potentially small effect sizes that may exist. As such, we also needed a more automated method to generate quantitative metrics for measuring civility across a larger set of comments.

We employed the Moral Foundation Dictionary version 2 (MFD-2) (Frimer et al. 2017) to measure the level of moral values present in a given comment. For each comment, this method outputs five values (what we call "MFD scores") between -1 and 1 across the five moral foundations: Care/Harm; Fairness/Cheating; Loyalty/Betrayal; Authority/Subversion; and Sanctity/Degradation. Negative values closer to -1 indicate that a

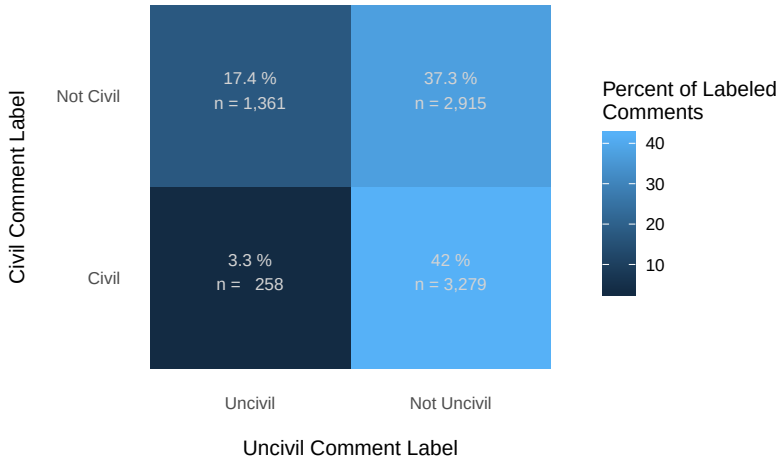

Figure 2: A confusion matrix showing the overlap of Civil and Uncivil labels for all comments labeled in this study. Nearly two-thirds of labeled comments were either only civil (42%) or only uncivil (17.4%), while over one-third of the comments were determined to contain neither (37.3%), and only rarely was a comment labeled with both civil and uncivil language (3.3%).

comment is more similar to the vice, and positive values closer to 1 indicate a comment is more similar to the virtue. We also computed the average across the five foundations for each comment and call this the "average MFD score." The supplementary material contains more technical details on how these scores were calculated.

Before looking into our hypotheses, analyses investigated whether there is any relationship between the civility labeling data and MFD scores. MFD scores were calculated for all comments that were reviewed and labeled by students. Figure 3 on the next page shows the distribution of the average MFD score for comments with civil and uncivil labels. Across all five foundations, comments labeled as civil tended to have higher MFD scores (closer to the virtue) than those labeled as not civil. The average MFD score for comments labeled civil was 0.053 compared to 0.032 for those labeled not civil (Cohen's d = 0.52). A Welch's two sample t-test confirms that the difference in mean is statistically significant ($p < 2^{e-16}$). Similarly, comments labeled uncivil tended to have lower MFD scores (closer to the vice) than comments labeled not uncivil. The average MFD score for comments labeled uncivil was 0.018 compared to 0.047 for comments labeled not uncivil (Cohen's d = 0.88. A Welch's two sample t test confirms that the difference in mean is statistically significant ($p < 2^{e-16}$). While the civility labeling can provide a more accurate insight into our specific definition of civility, the MFD approach appears to have enough overlap with our civility labeling to prove useful in analyzing much larger datasets.

4.4.3 User Reports

Nextdoor users can report other users' comments they deem offensive or otherwise inappropriate for the platform (Nextdoor 2022c). A user can choose to report a comment for any reason, though the platform provides tools for a user to indicate why they are reporting that comment. Comments that are reported are not necessarily uncivil, inappropriate, or otherwise offensive. Reported comments are sometimes reviewed and removed by other users in the neighborhood, while in some cases the platform reviews

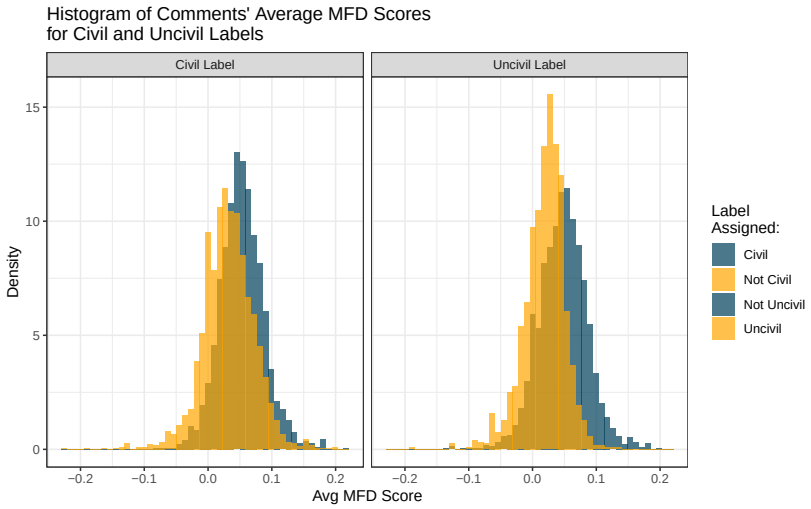

Figure 3: Distribution of the algorithmically calculated MFD scores (low scores indicate language associated with moral vice; high scores indicate language associated with moral virtues) on comments that were labeled using the civility labeling task. On the left, comments assigned a civil label (blue) had a higher average MFD of 0.053, compared to 0.032 for comments that did not have a civil label (yellow). On the right, comments assigned an uncivil label (yellow) had a lower average MFD of 0.018, compared to 0.047 for comments that did not have an uncivil label (blue).

and removes them. Our dataset from Nextdoor included information on whether a comment was reported, but not the result of the report (whether or not it was deemed to violate any platform rules).

Of the three types of metric used here, user reports are particularly important because they reflect some action users took. The first two indices used in this study (civility labels and MFD scores) are inferences made about the civility of the online discussion based on its content. By contrast, complaining to the platform through these reports is an action taken by a user to flag incivility. Unfortunately, there is no corresponding action that users can take to signal civil discussions.

4.5 Dataset

This study collected two different datasets from Nextdoor. The first, smaller, dataset was used exclusively to source comments for the civility labeling task to test *H1a* and *H2a*. This dataset was comprised of 100 Neighborhood Posts that were randomly sampled among all the Neighborhood Posts created in October 2020 within the study that became groups after prompting from Nextdoor. Half of these sampled Neighborhood Posts (50 of 100) were converted into a group that was assigned to the Guidelines condition, while the other 50 were assigned to the No Guidelines condition. This dataset also included the corresponding Group Comments.

The second dataset was a much larger dataset used for the more quantitative analyses. This dataset was used to calculate the MFD scores to test hypotheses *H1b* and *H2b* and to analyze the user reports to test hypotheses *H1c* and *H2c*. Table 2 displays the summary statistics for this dataset, which included all Neighborhood Posts involved in this study in May 2021 (n = 7,539) as well as the corresponding 1,118,031 Neighborhood Post Comments. The number of Neighborhood Post Comments per Neighborhood Post ranges from a minimum of 70 comments to a maximum of 1,141 comments on a single Neighborhood Post. Of the 7,539 Neighborhood Posts, 312 authors (4.1%) created a

group after being prompted. An analysis was conducted, included in the supplemental materials (S3), to look for the differences in the 4.1% of posts which chose to create a group and those that did not. The 312 Neighborhood Posts that became groups had a total of 5,345 Group Posts with a minimum of 2 and a maximum 237 Group Posts in any single group. These Group Posts had a total of 26,201 Group Comments with a minimum of 2 and a maximum 223 Group Comments on any single Group Post.

Table 2: Summary statistics of the data in the larger quantitative dataset. This dataset was used to calculate the MFD scores used for *H1b* and *H2b* and for analyses of user-submitted reports to test *H1c* and *H2c*.

Unit of Measurement	Amount
All Neighborhood Posts in Quantitative Dataset	
Neighborhood Posts	7,539
Neighborhood Post Comments	1,118,031
Neighborhood Post Comments per Neighborhood Post (Min-Max; [Mean])	70–1,141; [150]
Neighborhood Posts that Create Groups	
Neighborhood Posts	312
Neighborhood Post Comments	57,487
Neighborhood Post Comments per Neighborhood Post (Min-Max; [Mean])	73–1,048; [169]
Newly Formed Groups	
Groups	312
Group Posts	5,345
Group Posts per Group (Min-Max; [Mean])	2–237; [16]
Group Post Comments	26,201
Group Post Comments per Group Post (Min-Max; [Mean])	2–223; [7]

5 Results

5.1 Neighborhood Post Comments vs. Group Comments

To investigate what happens when a popular conversation moves into a group setting (*H1a/b/c*), this study analyzed Neighborhood Posts in which the author chose to create a group after being prompted to do so. We compared our three measures between the first 70 Neighborhood Post Comments—before any group had been created—to the Group Comments made within the group that was later created.

We randomly sampled and manually labeled 4,000 comments. First, 20 Neighborhood Post Comments were randomly sampled from among the first 70 Neighborhood Post Comments on 100 different Neighborhood Posts, totaling 2,000 labeled Neighborhood Post Comments made before the post author was asked to create a group. We then compared these comments to 20 randomly sampled Group Comments made from each of the 100 newly formed groups that were created, totaling 2,000 labeled Group Comments. This allowed us to compare the civility of comments made on the Neighborhood Posts before any intervention from Nextdoor (pre-test) to the civility of Group Comments made in the group created to discuss the same topic (post-test).

Using a chi-squared test, we found a significant difference in the proportion of both civil and uncivil comments. The proportion of comments classified as "civil" on the Neighborhood Post Comments was 0.41 compared to 0.56 for Group Comments ($p < 2^{e-16}$). Similarly, the proportion of comments classified as "uncivil" was 0.23 for Neighborhood Post Comments compared to 0.15 for Group Comments ($p = 1.3^{e-13}$) (Figure 4).

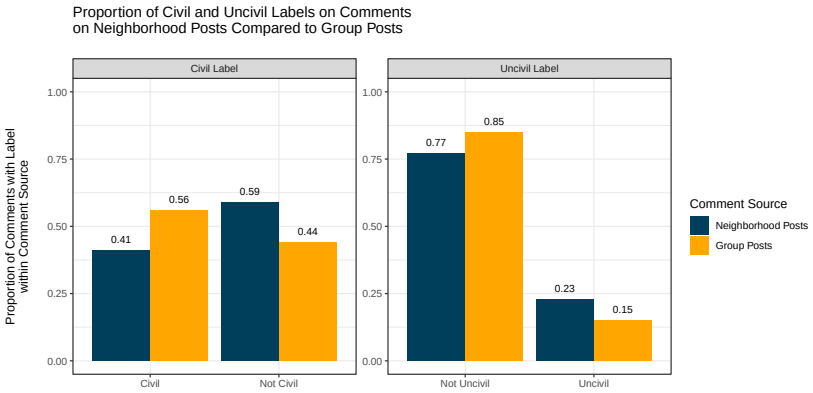

Figure 4: Proportion of comments' civil (left) and uncivil (right) classifications for Neighborhood Post Comments vs. Group Comments. Neighborhood Post Comments contain a lower proportion of comments labeled Civil and a higher proportion of comments labeled Uncivil compared to Group Comments.

A similar analysis was conducted using the much larger dataset for MFD scores and user reports. Again limiting ourselves to Neighborhood Posts that created groups after prompting from Nextdoor, we compared MFD scores on the first 70 Neighborhood Post Comments to scores on all Group Comments in the newly formed groups. Across all five foundations, we observed statistically significant increases in MFD scores for Group Comments compared to Neighborhood Post Comments (Table 3).

Table 3: Individual and average MFD scores associated with H1b

Moral Foundation	Group Comments	Neighborhood Post Comments	Cohen's d
Care	0.067	0.055	0.41
Fairness	0.023	0.019	0.24
Loyalty	0.056	0.048	0.37
Authority	0.090	0.084	0.27
Sanctity	0.042	0.037	0.23
Mean MFD	0.056	0.049	0.37

Note: The Cohen's d is the size of the difference in mean relative to the standard deviation of the data. A larger value indicates a stronger relative effect size. All differences in mean in the above tables are statistically significant under hypothesis testing.

Lastly, we used the larger dataset to analyze reports of comments made by users. While reporting comments is a relatively rare occurrence, given the number of comments in our dataset, we still observe meaningful differences. The proportion of Group Comments with one or more reports was 0.2%, while 1.4% of the first 70 Neighborhood Post Comments received one or more reports ($p < 2^{e-16}$). This indicates that Nextdoor members were significantly more likely to report comments made on Neighborhood Posts than Group Comments.

Across the three methods used—human labeling, MFD scores, and member reports—we

observed consistent results. Comments made within these groups were more likely to be labeled civil, less likely to be labeled uncivil, had relatively higher MFD scores, and were reported less often than comments made on the Neighborhood Posts. All of these results showed strong support for *H1a, H1b, and H1c*.

5.2 The Effect of Announcement of Group Guidelines on Civil Discussion

When Neighborhood Post authors chose to create a new group from their Neighborhood Post, that newly formed group was randomly assigned to either a Guidelines or No Guidelines condition in which new members joining this new group were shown or not shown a set of guidelines (Figure 1 on page 9). Comments made in these newly formed groups were sampled and labeled using the civility codebook. Twenty Group Comments were randomly sampled from each of 100 groups for a total of 2,000 labeled Group Comments. These 2,000 Group Comments were evenly split between groups in each condition (20 Group Comments from 50 Guidelines groups and 20 Group Comments from 50 No Guidelines groups). Comparing these two sets of Group Comments using a chi-squared test reveals a small and significant difference in the proportion of comments classified as civil; the No Guidelines group had a slightly higher proportion of Group Comments classified as civil. However, there was no statistically significant difference in the proportion of Group Comments labeled uncivil. The proportion of Group Comments with civil labels in Guidelines groups was 0.538 vs. 0.588 for No Guidelines groups (p = 0.026). The proportion of Group Comments classified as uncivil in Guidelines groups was 0.159 compared to 0.143 for Group Comments in No Guidelines groups (p = 0.3481) (Figure 5).

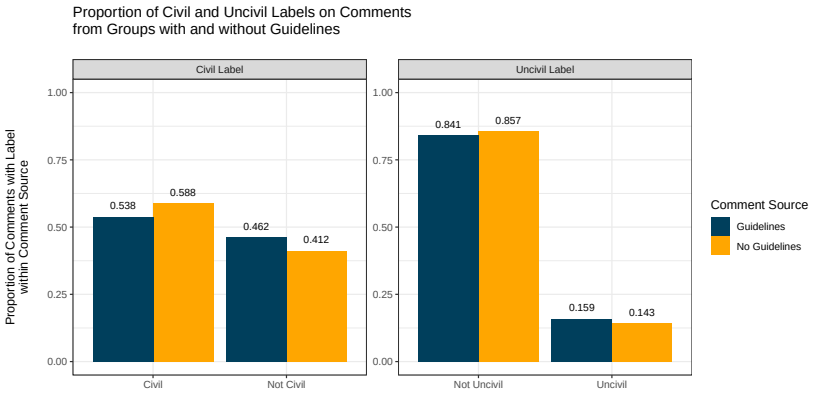

Figure 5: Proportion of comments' civil (left) and uncivil (right) classifications for Group Comments in Guidelines vs. No Guidelines groups. Guidelines Group Comments contain a very slightly lower proportion of comments labeled Civil and an equal proportion of comments labeled Uncivil compared to No Guidelines Group Comments.

Given the small effect on civility that was observed in the manually labeled data, the much larger dataset provides us an opportunity to more easily detect smaller effects that may result from the presentation of guidelines. Across all five moral foundations, there was a small but statistically significant increase in MFD scores for Group Comments in Guidelines groups compared to No Guidelines groups (Table 4 on the next page).

Lastly, we found that Group Comments in Guidelines groups were less likely to be reported by members than Group Comments in No Guidelines groups. The proportion of Guidelines groups' Group Comments that had one or more reports was 0.33%, compared to 0.72% for the Group Comments in No Guidelines groups ($p < 2^{e-16}$). This indi-

Table 4: Individual and Average MFD Scores Associated with H2b

Moral Foundation	Group Comments	Neighborhood Post Comments	Cohen's d
Care	0.033	0.031	0.02
Fairness	0.012	0.008	0.09
Loyalty	0.032	0.028	0.07
Authority	0.071	0.070	0.02
Sanctity	0.034	0.032	0.03
Mean MFD	0.036	0.034	0.05

Note: The Cohen's d is the size of the difference in mean relative to the standard deviation of the data. A larger value indicates a stronger relative effect size. All differences in mean in the above tables are statistically significant under hypothesis testing.

cates that Nextdoor members were significantly less likely to report Group Comments made in groups with guidelines than in those without guidelines.

Our second set of hypotheses examined the effect of showing or not showing guidelines to new group members on their participation in the newly formed groups. Here, the results are mixed. With the relatively small amount of human labeling, we saw a very small decrease in the proportion of civil comments but no difference in the proportion of uncivil comments in groups with guidelines compared to those without. Therefore, *H2a* was not supported. However, we did observe statistically significant changes in the MFD scores and member reports of Group Comments, which supports *H2b* and *H2c*. Overall, we found some support for *H2*: showing guidelines to new members does appear to have a positive, albeit small, impact on the conversations that follow in those groups.

6 Discussion

The nature of online platform content has become a widespread concern among policy makers, legal scholars, and the general public (Gorwa 2019; Klonick 2017). These concerns have led to normative questions about whether and how content should be managed alongside empirical research on how such management might be possible. Many platforms have utilized traditional legal frameworks to manage certain platform behaviors, which involve progressively severe sanctions ranging from takedowns to user suspensions and exclusions (Tyler et al. 2021). These efforts have been primarily directed at managing and regulating the amount of antisocial behavior and negative experiences. This focus on reducing negative or otherwise offensive content can often come at the cost of exploring how to foster more positive, pro-social content and connections on the platforms. This study works to fill that gap by promoting more positive, civil content in online interactions.

Our results demonstrate that more civil interactions among users can be encouraged by altering the design and architecture of the online environment within which the interaction occurs. The level of civility and moral values of Neighborhood Post Comments increased, while the number of reports of those comments made by users and the level of incivility decreased when conversations about a given topic were moved away from Neighborhood Posts and into new groups. Both architectural features—encouraging the formation of a group and proactively providing guidelines about civil interactions—were associated with improvements in discussions and behaviors. The findings complement previous research (Matias 2019; Tyler et al. 2021) which shows that announcing com-

munity guidelines or providing education not only decreases antisocial behaviors (as shown in previous research); it can also lead to more pro-social interactions among users. This research advances the existing literature by demonstrating the need for online environments that clearly communicate expectations and norms to users in order to encourage more pro-social interactions.

Regarding group formation, there have been general concerns about decreased membership in civic groups and civic engagement in recent decades (Putnam 2000). Although groups on social media may enhance online user engagement and offline participation (especially in politics) (Conroy, Feezell, and Guerrero 2012), these groups have been criticized for spreading misinformation and hate speech, as well as creating group polarization (Del Vicario et al. 2016; Merrill and Åkerlund 2018). However, our findings suggest that online groups might encourage users to participate in more civil interactions if appropriate guidelines are provided to new group members.

Although calls for civility have a strong moral appeal, we acknowledge that such an approach may be used to silence and harass feared or subordinate groups (Jamieson et al. 2017; Zurn 2013). Incivility might sometimes be beneficial in terms of draw attention and passionate engagement from others, or might even be required for some groups to get their point across (Cohen 1960). However, hate speech or uncivil discussion on social media platforms is often cited as a reason that many choose not to engage in discussions online (Kruse, Norris, and Flinchum 2017); minority groups may be more likely to be targeted with hate speech and incivility online (Vogels 2021). Moreover, compared to content moderation, such as removing uncivil content from platforms (which potentially suppresses free speech), encouraging civil behavior on social media platforms via indirect platform interventions would help promote individual and community well-being without silencing minority views.

This study makes theoretical contributions by providing empirical evidence of how platform architecture influences users' behaviors by analyzing a large volume of social media user data. In addition, future studies can use the codebook we developed for this study to operationalize and measure the concept of civility to examine how any number of platform architectures and strategies not only decrease uncivil conversations but also increase civil interactions. Finally, by collaborating with a social media platform, this study applied the theory of civility in a robust and practical setting.

However, this study is not without limitations. First, our analyses focused on conversations in which individuals chose to create groups from Neighborhood Posts. Since this was a relatively rare occurrence (only 4.1% of Neighborhood Posts resulted in a group being created), we cannot assume that all conversations on a given social media platform would be affected by similar interventions. The conversations that became groups appeared to, on average, start at a higher level of civility than the majority of conversations that did not become groups. On the one hand, this could suggest that such platform architecture interventions could have an even greater impact on the conversations that did not opt in to this intervention in our study. On the other hand, it could suggest that the civil conversations that self-selected into this group-forming intervention were more easily nudged towards even greater civility. As is the case with many interventions designed to shift community norms, there will never be a "silver bullet," and the two interventions presented here should be considered alongside other design patterns as platforms build their online environments.

A second limitation is that this study assumes that users' moral and pro-social beliefs are reflected in the language they use in social media posts and comments. Language has typically been regarded as the most common and reliable way for people to indicate their thoughts and emotional states, thereby reflecting who they are and their social re-

lationships (Tausczik and Pennebaker 2009). However, scholars have argued that social media users use different types of language based on their perceptions of a post's potential audience. Future studies could examine how social media users employ specific types of language to indicate their moral and pro-social beliefs, depending on the specific contexts. In addition, individual coders who conducted our civility labeling might have differing moral beliefs that affect their interpretation of civility in the comments they evaluated. To avoid this influence, we conducted multiple training sessions and tested inter-coder reliability. Nevertheless, it would be worth examining how individual coders' moral beliefs may affect their coding of the moral and pro-social beliefs reflected in the posts.

Another area for future inquiry is the exact pathway through which the group setting encourages more civil conversations. In this study, authors of popular posts were encouraged to create groups to facilitate a more focused discussion of a topic that was clearly resonating with their neighborhood. Our results indicate that these groups facilitated more civil conversations than the equivalent discussions that occurred at the neighborhood-wide level. The most basic distinction between these two settings (neighborhood-wide and group) is the size of the audience and the number of participants. However, groups also require those conversing to opt in to having a conversation about a given topic with others in the group. The group setting can also permit organizers to moderate some aspects of the discussion. Our second intervention illustrated that basic guidelines and ground rules for a conversation can be established. Further research should explore which factors within these group settings may contribute to the civility of conversations, and how.

Given that online engagement can be related to offline civic engagement (Putnam 2000), platform-level interventions might have other potential benefits to society. Future work could examine other consequences of group formation and announcing guidelines to new group members, including users' perceptions of or general attitudes toward the platform and their offline civic engagement or participation.

Finally, most previous efforts to measure the content of online interactions have focused on identifying negative content (e.g., hate speech, nudity). This study demonstrates that there are several viable mechanisms for capturing both positive and negative content, including creating a theory-based set of indicators of civility/incivility and drawing on existing dictionaries based on models of positive/negative words and phrases. This study found that these two approaches converged in their identification of both civil and uncivil content, suggesting that both are valid indicators of platform discussions.

References

Araque, Oscar, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. "MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction." *Knowledge-based systems* 191 (March): 105184. https://doi.org/10.1016/j.knosys.2019.105184. https://doi.org/10.1016/j.knosys.2019.105184.

Bender, Jacqueline L, Maria-Carolina Jimenez-Marroquin, and Alejandro R Jadad. 2011. "Seeking support on facebook: a content analysis of breast cancer groups." *Journal of medical Internet research* 13, no. 1 (February): e16. https://doi.org/10.2196/jmir.1560. https://doi.org/10.2196/jmir.1560.

Benn, Stanley I, and SI Benn. 1988. *A theory of freedom.* Cambridge University Press, August. https://doi.org/10.1017/cbo9780511609114. https://doi.org/10.1017/cbo9780511609114.

Blom, Robin, Serena Carpenter, Brian J Bowe, and Ryan Lange. 2014. "Frequent contributors within US newspaper comment forums: An examination of their civility and information value: An Examination of Their Civility and Information Value." *American Behavioral Scientist* 58, no. 10 (March): 1314–28. https://doi.org/10.1177/0002764214527094. https://doi.org/10.1177/0002764214527094.

Calhoun, Cheshire. 2000. "The virtue of civility." *Philosophy & public affairs* 29, no. 3 (July): 251–75. https://doi.org/10.1111/j.1088-4963.2000.00251.x. https://doi.org/10.1111/j.1088-4963.2000.00251.x.

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, no. 4 (June): 658–79. https://doi.org/10.1111/jcom.12104. https://doi.org/10.1111/jcom.12104.

Cohen, Jacob. 1960. "A coefficient of agreement for nominal scales." *Educational and psychological measurement* 20, no. 1 (April): 37–46. https://doi.org/10.1177/001316446002000104. https://doi.org/10.1177/001316446002000104.

Conroy, Meredith, Jessica T. Feezell, and Mario Guerrero. 2012. "Facebook and political engagement: A study of online political group membership and offline political engagement." *Computers in Human Behavior* 28, no. 5 (September): 1535–46. https://doi.org/10.1016/j.chb.2012.03.012. https://doi.org/10.1016/j.chb.2012.03.012.

Del Vicario, Michela, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2016. "Echo chambers: Emotional contagion and group polarization on facebook." *Scientific reports* 6, no. 1 (December): 1–12. https://doi.org/10.1038/srep37825. https://doi.org/10.1038/srep37825.

Frimer, J, J Haidt, J Graham, M Dehghani, and R Boghrati. 2017. "Moral foundations dictionaries for linguistic analyses, 2.0." *Unpublished Manuscript. Retrieved from: www. jeremyfrimer. com/uploads/2/1/2/7/21278832/summary. pdf.*

Gahagan, Kassandra, J Mitchell Vaterlaus, and Libby R Frost. 2016. "College student cyberbullying on social networking sites: Conceptualization, prevalence, and perceived bystander responsibility." *Computers in human behavior* 55 (February): 1097–105. https://doi.org/10.1016/j.chb.2015.11.019. https://doi.org/10.1016/j.chb.2015.11.019.

Gorwa, Robert. 2019. "What is platform governance?" *Information, Communication & Society* 22, no. 6 (February): 854–71. https://doi.org/10.1080/1369118x.2019.1573914. https://doi.org/10.1080/1369118x.2019.1573914.

Graham, Jesse, and Jonathan Haidt. 2012. "Sacred values and evil adversaries: A moral foundations approach.," https://doi.org/10.1037/13091-001.

Graham, Jesse, Jonathan Haidt, and Brian A Nosek. 2009. "Liberals and conservatives rely on different sets of moral foundations." *Journal of personality and social psychology* 96, no. 5 (May): 1029. https://doi.org/10.1037/a0015141. https://doi.org/10.1037/a0015141.

Grover, Ted, Elvan Bayraktaroglu, Gloria Mark, and Eugenia Ha Rim Rho. 2019. "Moral and affective differences in us immigration policy debate on twitter." *Computer Supported Cooperative Work (CSCW)* 28, no. 3 (May): 317–55. https://doi.org/10.1007/s10606-019-09357-w. https://doi.org/10.1007/s10606-019-09357-w.

Haidt, Jonathan. 2008. "Morality." PMID: 26158671, *Perspectives on Psychological Science* 3, no. 1 (January): 65–72. https://doi.org/10.1111/j.1745-6916.2008.00063.x. eprint: https://doi.org/10.1111/j.1745-6916.2008.00063.x. https://doi.org/10.1111/j.1745-6916.2008.00063.x.

Haidt, Jonathan, and Jesse Graham. 2007. "When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize." *Social Justice Research* 20, no. 1 (May): 98–116. https://doi.org/10.1007/s11211-007-0034-z. https://doi.org/10.1007/s11211-007-0034-z.

Haidt, Jonathan, Jesse Graham, and Craig Joseph. 2009. "Above and below left–right: Ideological narratives and moral foundations." *Psychological Inquiry* 20, nos. 2-3 (August): 110–19. https://doi.org/10.1080/10478400903028573. https://doi.org/10.1080/10478400903028573.

Haidt, Jonathan, and Craig Joseph. 2004. "Intuitive ethics: How innately prepared intuitions generate culturally variable virtues." *Daedalus* 133, no. 4 (September): 55–66. https://doi.org/10.1162/0011526042365555. https://doi.org/10.1162/0011526042365555.

Han, Soo-Hye, LeAnn M Brazeal, and Natalie Pennington. 2018. "Is civility contagious? Examining the impact of modeling in online political discussions." *Social Media+ Society* 4, no. 3 (July): 2056305118793404. https://doi.org/10.1177/2056305118793404. https://doi.org/10.1177/2056305118793404.

Hille, Sanne, and Piet Bakker. 2014. "Engaging the Social News User: Comments on news sites and Facebook." *Journalism Practice* 8, no. 5 (April): 563–72. https://doi.org/10.1080/17512786.2014.899758. https://doi.org/10.1080/17512786.2014.899758.

Hoover, Joe, Mohammad Atari, Aida Mostafazadeh Davani, Brendan Kennedy, Gwenyth Portillo-Wightman, Leigh Yeh, and Morteza Dehghani. 2021. "Investigating the role of group-based morality in extreme behavioral expressions of prejudice." *Nature Communications* 12, no. 1 (July): 1–13. https://doi.org/10.1038/s41467-021-24786-2. https://doi.org/10.1038/s41467-021-24786-2.

Jaidka, Kokil, Alvin Zhou, and Yphtach Lelkes. 2019. "Brevity is the soul of Twitter: The constraint affordance and political discussion." *Journal of Communication* 69, no. 4 (July): 345–72. https://doi.org/10.1093/joc/jqz023. https://doi.org/10.1093/joc/jqz023.

Jamieson, Kathleen Hall, Allyson Volinsky, Ilana Weitz, and Kate Kenski. 2017. "The political uses and abuses of civility and incivility." *The Oxford handbook of political communication,* 205–18.

Jhaver, Shagun, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. ""Did You Suspect the Post Would be Removed?" Understanding User Reactions to Content Removals on Reddit: Understanding User Reactions to Content Removals on Reddit." *Proceedings of the ACM on human-computer interaction* 3, no. CSCW (November): 1–33. https://doi.org/10.1145/3359294. https://doi.org/10.1145/3359294.

Jin, Chang-Hyun. 2015. "The role of Facebook users' self-systems in generating social relationships and social capital effects." *New Media & Society* 17, no. 4 (October): 501–19. https://doi.org/10.1177/1461444813506977. https://doi.org/10.1177/1461444813506977.

Katsaros, Matthew, Kathy Yang, and Lauren Fratamico. 2022. "Reconsidering Tweets: Intervening During Tweet Creation Decreases Offensive Content." In *Proceedings of the International AAAI Conference on Web and Social Media,* 16:477–87.

Klonick, Kate. 2017. "The new governors: The people, rules, and processes governing online speech." *Harv. L. Rev.* 131:1598. https://doi.org/10.2139/ssrn.3332530. https://doi.org/10.2139/ssrn.3332530.

Kruse, Lisa M., Dawn R. Norris, and Jonathan R. Flinchum. 2017. "Social Media as a Public Sphere? Politics on Social Media." *The Sociological Quarterly* 59, no. 1 (October): 62–84. https://doi.org/10.1080/00380253.2017.1383143. https://doi.org/10.1080/00380253.2017.1383143.

Ksiazek, Thomas B. 2015. "Civil interactivity: How news organizations' commenting policies explain civility and hostility in user comments." *Journal of Broadcasting & Electronic Media* 59, no. 4 (October): 556–73. https://doi.org/10.1080/08838151.2015.1093487. https://doi.org/10.1080/08838151.2015.1093487.

Lampe, Cliff, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. 2014. "Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums." *Government Information Quarterly* 31, no. 2 (April): 317–26. https://doi.org/10.1016/j.giq.2013.11.005. https://doi.org/10.1016/j.giq.2013.11.005.

Matias, J Nathan. 2019. "Preventing harassment and increasing group participation through social norms in 2,190 online science discussions." *Proceedings of the National Academy of Sciences* 116, no. 20 (April): 9785–89. https://doi.org/10.1073/pnas.1813486116. https://doi.org/10.1073/pnas.1813486116.

Merrill, Samuel, and Mathilda Åkerlund. 2018. "Standing Up for Sweden? The Racist Discourses, Architectures and Affordances of an Anti-Immigration Facebook Group." *Journal of Computer-Mediated Communication* 23, no. 6 (September): 332–53. https://doi.org/10.1093/jcmc/zmy018. https://doi.org/10.1093/jcmc/zmy018.

Meta Platforms, Inc. 2022. "Meta Company Info." Accessed August 30, 2022. https://about.facebook.com/company-info/.

Nextdoor. 2022a. "About Nextdoor, the Neighborhood Network." Accessed August 30, 2022. https://about.nextdoor.com/.

———. 2022b. "How to create and edit a group." Accessed August 30, 2022. https://help.nextdoor.com/s/article/How-to-create-a-group?language=en_US.

———. 2022c. "How to report a post." Accessed August 30, 2022. https://help.nextdoor.com/s/article/How-to-report-content?language=en_US.

Nilsson, Artur, Arvid Erlandsson, and Daniel Västfjäll. 2016. "The congruency between moral foundations and intentions to donate, self-reported donations, and actual donations to charity." *Journal of Research in Personality* 65 (December): 22–29. https://doi.org/10.1016/j.jrp.2016.07.001. https://doi.org/10.1016/j.jrp.2016.07.001.

Papacharissi, Zizi. 2004. "Democracy online: Civility, politeness, and the democratic potential of online political discussion groups." *New media & society* 6, no. 2 (April): 259–83. https://doi.org/10.1177/1461444804041444. https://doi.org/10.1177/1461444804041444.

Phua, Joe, Seunga Venus Jin, and Jihoon Jay Kim. 2017. "Uses and gratifications of social networking sites for bridging and bonding social capital: A comparison of Facebook, Twitter, Instagram, and Snapchat." *Computers in human behavior* 72 (July): 115–22. https://doi.org/10.1016/j.chb.2017.02.041. https://doi.org/10.1016/j.chb.2017.02.041.

Putnam, Robert D. 2000. "Bowling Alone: America's Declining Social Capital." In *Culture and Politics,* 223–34. Palgrave Macmillan US. https://doi.org/10.1007/978-1-349-62397-6_12. https://doi.org/10.1007/978-1-349-62397-6_12.

Rawls, John. 1996. *Political liberalism: The John Dewey essays in philosophy.* New York : Columbia University Press.

Reader, Bill. 2012. "Free press vs. free speech? The rhetoric of "civility" in regard to anonymous online comments." *Journalism & Mass Communication Quarterly* 89, no. 3 (May): 495–513. https://doi.org/10.1177/1077699012447923. https://doi.org/10.1177/1077699012447923.

Ribeiro, Manoel Horta, Justin Cheng, and Robert West. 2022. "Post Approvals in Online Communities." In *Proceedings of the International AAAI Conference on Web and Social Media,* 16:335–46.

Rieger, Diana, Anna Sophie Kümpel, Maximilian Wich, Toni Kiening, and Georg Groh. 2021. "Assessing the Extent and Types of Hate Speech in Fringe Communities: A Case Study of Alt-Right Communities on 8chan, 4chan, and Reddit." *Social Media+ Society* 7 (4): 20563051211052906.

Rowe, Ian. 2014. "Civility 2.0: A comparative analysis of incivility in online political discussion." *Information, communication & society* 18, no. 2 (July): 121–38. https://doi.org/10.1080/1369118x.2014.940365. https://doi.org/10.1080/1369118x.2014.940365.

Ruiz, Carlos, David Domingo, Josep Lluís Micó, Javier Díaz-Noci, Koldo Meso, and Pere Masip. 2011. "Public sphere 2.0? The democratic qualities of citizen debates in online newspapers." *The International journal of press/politics* 16, no. 4 (September): 463–87. https://doi.org/10.1177/1940161211415849. https://doi.org/10.1177/1940161211415849.

Santana, Arthur D. 2013. "Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards." *Journalism practice* 8, no. 1 (July): 18–33. https://doi.org/10.1080/17512786.2013.813194. https://doi.org/10.1080/17512786.2013.813194.

Seering, Joseph, Tianmi Fang, Luca Damasco, Mianhong 'Cherie' Chen, Likang Sun, and Geoff Kaufman. 2019. "Designing user interface elements to improve the quality and civility of discourse in online commenting behaviors." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* 1–14. ACM, May. https://doi.org/10.1145/3290605.3300836. https://doi.org/10.1145/3290605.3300836.

Smith, Elizabeth S, and Alison Bressler. 2013. "Who taught you to talk like that?: The university and online political discourse." *Journal of Political Science Education* 9, no. 4 (October): 453–73. https://doi.org/10.1080/15512169.2013.835565. https://doi.org/10.1080/15512169.2013.835565.

Tausczik, Yla R., and James W. Pennebaker. 2009. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29, no. 1 (December): 24–54. https://doi.org/10.1177/0261927x09351676. https://doi.org/10.1177/0261927x09351676.

Twitter, Inc. 2022. "Twitter - About Our Company." Accessed August 30, 2022. https://about.twitter.com/en/who-we-are/our-company.

Tyler, Tom, Matt Katsaros, Tracey Meares, and Sudhir Venkatesh. 2021. "Social media governance: can social media companies motivate voluntary rule following behavior among their users?" *Journal of experimental criminology* 17, no. 1 (December): 109–27. https://doi.org/10.1007/s11292-019-09392-z. https://doi.org/10.1007/s11292-019-09392-z.

Tyler, Tom R, Jonathan Jackson, and B Bradford. 2014. "Psychology of procedural justice and cooperation."

Vogels, Emily A. 2021. "The State of Online Harassment." *Pew Research Center,* https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/.

Waldron, Jeremy. 2013. "Civility and formality." *NYU School of Law, Public Law Research Paper,* nos. 13-57, https://doi.org/10.2139/ssrn.2326759. https://doi.org/10.2139/ssrn.2326759.

Walter, Annemarie, and Keena Lipsitz. 2021. "How Moral Value Commitments Shape Responses to Political Civility and Incivility." *American Politics Research* 49, no. 4 (March): 359–67. https://doi.org/10.1177/1532673x211004137. https://doi.org/10.1177/1532673x211004137.

Welsch, Heinz. 2020. "Moral foundations and voluntary public good provision: the case of climate change." *Ecological Economics* 175 (September): 106696. https://doi.org/10.1016/j.ecolecon.2020.106696. https://doi.org/10.1016/j.ecolecon.2020.106696.

Whitman, James Q. 2000. "Enforcing Civility and Respect: Three Societies." *The Yale Law Journal* 109, no. 6 (April): 1279. https://doi.org/10.2307/797466. https://doi.org/10.2307/797466.

YouTube. 2022. "About YouTube." Accessed August 30, 2022. https://about.youtube/.

Zurn, Christopher F. 2013. "Political civility: Another illusionistic ideal." *Public Affairs Quarterly* 27 (4): 341–68. https://www.jstor.org/stable/43575586.

Authors

Jisu Kim is an Assistant Professor at the Singapore Institute of Technology, where she teaches in the Digital Communications and Interactive Media program.

Curtis McDonald is a Ph.D. student at Yale University's Department of Statistics and Data Science.

Paul Meosky is a J.D. student at Yale Law School.

Matthew Katsaros is the Director of the Social Media Governance Initiative at the Justice Collaboratory at Yale Law School.

Tom Tyler is the Macklin Fleming Professor of Law at Yale University and a Professor in the Yale Psychology Department.

Acknowledgements

The authors would like to thank Farzaneh Badiei for contributions to this work. They would also like to thank the team at Nextdoor for their collaboration on this project, including Nasim Farsinia, Kyle Miller, Shane Butler, and Zach Kahn. We owe a great debt of gratitude to all of the research assistants who assisted with civility labeling.

Data Availability Statement

Data for this study is not available for replication, as it is comprised of user-generated comments discussing local neighborhood issues that regularly contain personally identifiable information of people and places, which makes it difficult to anonymize (a necessary step to make data available for replication).

Funding Statement

This study was funded and supported by the Justice Collaboratory at Yale Law School. One of the co-authors (Matt Katsaros) worked briefly with Nextdoor as a paid research advisor for 1 month approximately 1 year prior to working on this study.

Ethical Standards

The study was reviewed and approved by the Yale University IRB. Approval was given June 25, 2020 (approval number 2000028519).

Keywords

Civility; morality; moral foundation dictionary (MFD); online discussion; procedural justice; social media.