Journal of Computer-Mediated Communication, 2023, zmad018
https://doi.org/10.1093/jcmc/zmad018
Full-length Research Article

International Communication Association
OXFORD

Differential perceptions of and reactions to incivil and intolerant user comments

Anna Sophie Kümpel [1,*], Julian Unkel [2]

[1]Institute of Media and Communication, TU Dresden, Dresden, Saxony, Germany
[2]Department of Media and Communication, LMU Munich, Munich, Bavaria, Germany

*Corresponding author: Anna Sophie Kümpel. Email: anna.kuempel@tu-dresden.de

Abstract

Building on recent research that challenges the notion that norm violations in online discussions are inherently detrimental, this study relies on a distinction between *incivil* and *intolerant* user comments and investigates how online users perceive and react to these distinct forms of antinormative discourse online. Conducting a preregistered factorial survey experiment with a nationally representative sample of $n = 964$ German online users, we presented participants with manipulated user comments that included statements associated with incivil (profanity; attacks toward arguments) and intolerant discourse (offensive stereotyping; violent threats). The results show that intolerant statements consistently lead to higher perceptions of offensiveness and harm to society as well as an increased intention to delete the comment containing the statement, whereas incivil statements do not. An exploratory multiverse analysis further suggests that these effects remain robust across a variety of analytical decisions.

Lay Summary

Online discussions often violate social norms. People can be rude, use offensive language, or even make threats. But research shows that not all norm violations are equal. *Incivility* (e.g., using profanity) is primarily a way of getting attention and can still allow for meaningful discussions. *Intolerance* (e.g., using violent threats), on the other hand, is meant to cause harm and goes against what is accepted in a democratic society. Our study asks whether online users react differently to incivil and intolerant statements in user comments. The results show that users find intolerant comments worse and consider them more harmful than incivil comments. These findings are consistent even when looking at the data in different ways.

Keywords: incivility, intolerance, online discussions, user comments, social media, experimental research, multiverse analysis

Citizen now routinely use news websites or social media platforms to discuss societal issues and engage in informal political talk (Pennington & Winfrey, 2021; Ziegele et al., 2018). While from a normative standpoint, online discussions should represent diverse viewpoints and be respectful and rational, research has repeatedly highlighted the pervasiveness of *incivility*: In diverse settings from political discussions on Twitter (e.g., Theocharis et al., 2020; Unkel & Kümpel, 2022) to news providers' websites or their outlets on Facebook (e.g., Rossini, 2022; Su et al., 2018), a considerable amount of user comments seems to contain norm-violating utterances. Although a growing body of literature has addressed this issue, a key problem is that operationalizations of incivility are diverse and inconsistent, making comparisons difficult (see also Chen et al., 2019). Moreover, many studies have failed to distinguish between impolite and interpersonal disrespectful utterances and those that threaten, damage, or breach democratic values (Oh et al., 2021; Rossini, 2019, 2022). This distinction, however, seems crucial when considering the potential effects of antinormative discourse. While *incivil* user comments—defined here in line with Rossini (2019, 2022) as comments featuring rude, harsh, or vulgar expressions—can serve strategic goals in discussions and are not necessarily detrimental to the quality of debates, *intolerant* user comments are those that jeopardize online discourse and, ultimately, social cohesion. Intolerant utterances deny others of an equal status and express a harmful or discriminatory intent toward individuals or groups based on their social identities, preferences, or beliefs (Rossini, 2019, 2022). Content-analytical studies have shown that intolerance and incivility occur in different discussion contexts and "that the core problem of uncivil society is intolerance, not incivility" (Oh et al., 2021, p. 104; see also Rossini, 2019, 2022), further pointing to the need to clearly differentiate between the two types of norm violations.

This content-analytical research, while essential in terms of conceptualizing and understanding the prevalence and different functions of incivility/intolerance, is based on message characteristics. However, we still know little about users' *perceptions* of and *reactions* to these types of antinormative discourse. Following calls for further experimental research by Rossini (2022, p. 417) and Muddiman (2019, p. 14), the present study investigates how distinct subtypes of incivility (profanity; attacks toward arguments) and intolerance (offensive stereotyping; violent threats) in user comments influence online users' perceptions regarding the comment's offensiveness and harm to society as well as their intention to delete the comment. Understanding users' perceptions of these different types of norm violations is not only important for developing effective moderation policies, but also for promoting a culture of respectful discourse, and encouraging responsible behavior in CMC environments. Relying on a preregistered online factorial survey experiment with a nationally representative sample of $n = 964$ German participants, our results show that

Associate Editor: Cuihua Shen
Received: 11 November 2022. **Revised:** 6 March 2023. **Accepted:** 28 April 2023

intolerant statements consistently lead to higher perceptions of offensiveness and harm to society as well as a higher likelihood to report the intention to delete the comment, whereas incivil statements do not. An exploratory multiverse analysis shows that these effects remain robust across a variety of analytical decisions, including different model specifications and controlling for theoretically relevant context- and person-specific characteristics.

Literature review

Incivility versus intolerance: differentiating antinormative discourse

In the light of a perceived rise of polarization and complaints about the quality of online discussions, incivility has become a "concept *du jour*" (Chen et al., 2019) in recent years, especially in research about social media (see also Rossini, 2019; Su et al., 2018). Various CMC theories—such as social presence theory, the reduced social cue perspective, or the social identity model of deindividuation effects—suggest that online information environments are particularly prone to encourage incivil communication (for an overview, see Kim, 2022). Due to a lower sense of social presence, the lack of nonverbal cues such as tone of voice or facial expressions, and processes of deindividuation (which lessen personal accountability and heighten conformity to group norms), people tend to be "more outrageous, obnoxious, or hateful in what they say" (Brown, 2018, p. 298) in CMC settings. On social media specifically, algorithmic amplification further boosts the visibility of incivil comments, which might then lead to even more incivility (Unkel & Kümpel, 2022).

Despite the increase in scholarly activities, incivility is still considered difficult to define—as evidenced both by respective claims by researchers (e.g., Rega & Marchetti, 2021; Theocharis et al., 2020) and the plethora of existing operationalizations in the literature (for an overview, see Bormann et al., 2022). However, there is some consensus that incivility is seen as a *violation of norms*. While recent research has identified a whole set of norms that incivil discourse violates (Bormann et al., 2022), at least two perspectives can be distinguished (Chen et al., 2019; Muddiman, 2019; Papacharissi, 2004; Rossini, 2019): (a) incivility as a violation of *politeness norms* and (b) incivility as a violation of *democratic norms*.

The first perspective, which is based on politeness theories (e.g., Fraser, 1990), regards vulgar or rude remarks, personal attacks, or the use of disrespectful language as incivil. By equating incivility with impoliteness, scholars rooted in this perspective have mainly focused on interpersonal interaction norms or—as Muddiman (2017) calls it—personal-level incivility. The second perspective, which is based on deliberative theories (e.g., Gutmann & Thompson, 1996), regards threats against democracy, stereotyping of marginalized groups, or discrimination as incivil. Labeled as public-level incivility by Muddiman (2017), this type of incivility has typically been conceptualized as more severe and is seen as conveying "disrespect for the collective traditions of democracy" (Papacharissi, 2004, p. 267). While some authors avoid the term "incivility" when referring to the violation of politeness norms and label respective utterances as impoliteness (e.g., Kalch & Naab, 2018; Papacharissi, 2004), others argue that this is the essence of incivility (e.g., Oh et al., 2021; Rossini, 2019, 2022) and that violations of democratic norms are better summarized under the term *intolerance* (Oh et al., 2021; Rossini, 2019, 2022).

We share the second view in this article and draw on Rossini's (2019, 2022) distinction between *incivility* and *intolerance*. Approaching (in)civility as a communicative practice, she conceptualizes incivility as a context-dependent feature of discourse that is characterized by the violation of discussion norms. Thus, incivil comments are defined as comments featuring expressions with a rude, disrespectful, or dismissive tone directed at other participants, their arguments, or at the discussion topic. Rossini (2022, p. 411) differentiates four subtypes of incivility, namely the use of profane or vulgar language, personal attacks, aspersions, and attacks toward arguments. For this study, we are focusing on the first and last subtype and investigate how the use of *profanity* (e.g., typical four-letter words such as "shit" or "fuck") and *attacks toward arguments* (i.e., dismissing or disqualifying a position) influence users' perceptions and reactions. This restriction was made in consideration of the pervasiveness of these subtypes in online discussions as well as necessary methodological considerations.[1] Importantly, being a communicative practice, incivility is first and foremost a *rhetorical tool* that people use to express their opinions or lend emphasis to their statements (Herbst, 2010; Rossini, 2019). Research has shown that it can increase attention or one's awareness of opposing viewpoints, and even foster political participation (Borah, 2014; Lee et al., 2022; Mutz, 2015). Therefore, the use of vulgar language or swearing may actually be contributing to forming opinions and understanding other people's positions. Especially propositional swearing (i.e., swearing that is used with intention)—the most likely form in text-based CMC—is often used to convey emotion or stress a point in a discussion (Jay & Janschewitz, 2008).

In contrast to incivility, intolerance is conceptualized not as a matter of tone, but of substance: a set of behaviors that are threatening to the values of democratic pluralism (Oh et al., 2021; Rossini, 2022). Accordingly, we define intolerant comments as expressing a harmful or discriminatory intent toward individuals or groups based on their identities, preferences, or beliefs. Thus, they are closely connected to the concept of *hate speech*, which can be seen as a subtype of intolerance that incites hatred or violence of people based on their belonging to a social group (Kunst et al., 2021; Schmid et al., 2022). However, in line with Rossini (2022, pp. 404–405), we rely on the broader concept of intolerance here, as it is not only restricted to exclusionary language focused on group-defining characteristics, but also encompasses other violations of democratic values or moral respect. Overall, Rossini (2022, p. 411) differentiates eight subtypes of intolerance, ranging from various forms of intolerance (e.g., toward political positions, sexual or religious freedom) to instances of racism, offensive stereotyping, and violent threats. Again, we are focusing on two subtypes in our study, namely *offensive stereotyping* (i.e., highlighting personal or cultural features in offensive ways) and *violent threats* (i.e., threatening the use of violence against persons/groups), as these can be clearly distinguished conceptually and are not limited to certain target populations or topics. Moreover, in a content analysis of Brazil's biggest online news outlet Portal UOL, these two subtypes were among the most prevalent forms of intolerance posted in response to political stories on both the outlet's website and Facebook page (Rossini, 2019, p. 149). In contrast to incivility, intolerance is less context-dependent, as it inevitably offends or undermines other people and thus signals a fundamental lack of mutual respect (Oh et al., 2021; Rossini, 2022).

Content-analytical research indicates that intolerance and incivility occur in noticeably different contexts (Oh et al., 2021; Rossini, 2019, 2022): Intolerance is more likely in discussions around minorities and policy-related topics—settings, where violations of democratic norms might be particularly severe. Incivility, on the other hand, is often associated with meaningful engagement such as calling attention to injustice or encouraging civic engagement. In terms of co-occurrence, research shows that about 11% (Rossini, 2022) to 12% (Oh et al., 2021) of incivil comments entail intolerance, while 36% (Oh et al., 2021) to 52% (Rossini, 2022) of intolerant comments also feature incivility. This suggests that intolerant comments are more likely to be both intolerant *and* incivil, but not the other way around. While these studies provide important insights into the prevalence and different functions of incivility and intolerance, they are based on message characteristics. As such, it remains largely unclear how users *perceive* these different forms of antinormative discourse.

Perceptions of and reactions to incivility and intolerance

Studies have repeatedly shown that it is "very much in the eye of the beholder" (Herbst, 2010, p. 3) whether comments are perceived as violating norms (Kenski et al., 2019, 2020; Muddiman, 2019; Stryker et al., 2016, 2022), and, as a result, whether users feel offended by such comments, discern them as harmful to society, or believe that corrective actions are needed. Relying on a national, diverse sample of U.S. respondents of voting age, Stryker et al. (2022) investigate how 20 types of "potential incivility" (p. 168) are evaluated. Focusing on the dimensions resembling the subtypes of incivility (profanity; attacks toward arguments) and intolerance (violent threats; offensive stereotyping) we are interested in, the results show that intolerant utterances—threatening or encouraging harm and using racial or sexual slurs—are perceived markedly worse than vulgarity or attacking one's stand on issues (i.e., incivil utterances). However, a problem with this study is that perceptions were only assessed on a 4-point scale ranging from "not at all uncivil" to "very uncivil," and that the different communicative behaviors were merely described instead of being presented as actual comments. Nevertheless, the results of earlier (experimental) survey research provide initial evidence that incivil and intolerant comments are not only conceptually different, but also perceived differently by individuals (Kenski et al., 2019, 2020; Muddiman, 2019; Stryker et al., 2016, 2022). In the context of our study, we are interested in two related, but distinct evaluations: perceptions of *offensiveness* and perceptions of *harm to society*. This allows us to gain insights into both users' more personal feelings and the emotional impact of incivility/intolerance as well as more socially oriented considerations resulting from norm-violating comments in online environments.

Individuals' perceptions are closely connected to their reactions and behavioral intentions. Past research has focused on a variety of reactions to incivil/intolerant comments—often connected to specific features provided in CMC environments—such as replying to comments, disliking, or reporting them (e.g., Gagrčin, 2022; Kalch & Naab, 2018; Kunst et al., 2021; Naab et al., 2021; Ziegele et al., 2020). As Bormann and colleagues (2022) argue, such reactions toward norm violations can be understood as a form of "explicit disapproval" (p. 349). Just like perceptions, this disapproval is not universal, but differs between communication norms, situations, and not least individuals. In the context of our study, we are interested in users' *deletion intention*, operationalized as the decision to remove a specific comment from the discussion. This is an obvious kind of disapproval, as the commenter is denied the right to participate in the conversation. Research on flagging (i.e., notifying platform providers of alleged norm violations)—which can be seen as a precursor to deletions—suggests that hateful comments are more likely to be flagged than "mere" disparaging comments (Kunst et al., 2021) and that direct calls for violence are most likely to be reported (Wilhelm et al., 2020). However, it should be noted that comments are also flagged for reasons other than norm violations, such as the use of partisan language (Muddiman & Stroud, 2017), pointing to the need to consider further context- and person-specific factors (see the next section).

Taken together, theoretical assumptions and content-analytical research suggest that intolerant utterances (in our case: *offensive stereotyping* and *violent threats*) are "the true democratic problem" (Oh et al., 2021, p. 105; see also Rossini, 2022), while incivil utterances (in our case: *profanity* and *attacks toward arguments*) might not only be less problematic, but—at least in certain situations—can even have positive effects for societal discourse. Building on the evidence provided above, we want to investigate how posts containing the mentioned subtypes of intolerance/incivility affect users' perceptions of offensiveness, their perceptions of the post's harm to society, as well as their intention to delete the post. In a first step, we are hypothesizing that posts that feature *any* kind of antinormative discourse will be perceived as more offensive, more harmful, and more in need of deletion than posts not containing these utterances:

> Posts containing profanity (**H1**) | attacks toward arguments (**H2**) | offensive stereotyping (**H3**) | violent threats (**H4**) will lead to (a) higher perceptions of offense, (b) higher perceptions of harm to society, and (c) higher deletion intentions than posts that do not contain the respective norm violation.

However, in direct comparison and in line with our reasoning, posts featuring intolerance should have *stronger* effects on perceptions and reactions than posts featuring incivility. Accordingly, we hypothesize:

> **H5:** The effects on (a) perceptions of offense, (b) perceptions of harm to society, and (c) deletion intentions of post features associated with intolerance (offensive stereotyping, violent threats) will be stronger than the effects of post features associated with incivility (profanity, attacks toward arguments).

The influence of context- and person-specific factors

Perceptions of and reactions to incivil and intolerant user comments may be influenced not only just by their "degree" of norm violations, but also by context-specific factors that relate, for example, to users' *attitudes toward the topic* in question (e.g., Borah, 2014; Kalch & Naab, 2018; Kim & Park, 2019; Muddiman, 2019). More specifically, users who have strong opinions about the subject matter should perceive an antinormative comment that (implicitly) criticizes their stance more negatively. Closely related to this, users' *perceived similarity to the commenting person* (e.g., regarding social status, partisan identity) as well as their *perceived similarity to the person/group targeted by the norm violation* should influence how offended they feel, how harmful they perceive the post to be, and their intention to delete the post. In congruence with theories of motivated reasoning (Kunda, 1990), users are likely to respond differently to incivility/intolerance expressed by people perceived to be in their own social group versus in opposing groups (Chen & Lu, 2017; Kim & Park, 2019; Muddiman, 2019). Likewise, if a post directly attacks a specific social group and the user is part of that group, intolerant utterances in particular might be perceived as more severe (Costello et al., 2019; Schmid et al., 2022; Williams et al., 2016).

In addition to these context-specific factors, which are directly tied to the content of the comment, more stable person-specific factors could be relevant as well. In political discussions, characteristics such as users' position on the *left–right spectrum*, their general *satisfaction with the political system*, or their *support for free speech* could be associated with how they perceive and react to comments featuring intolerance and/or incivility (Costello et al., 2019; Kenski et al., 2020; Riedl et al., 2021). Moreover, research has repeatedly shown that users' overarching *sociodemographic characteristics* (especially age, gender, and education), *personality traits*, and *(social) media use habits* are associated with how sensitive they are to incivility and intolerance (Bormann, 2022; Costello et al., 2019; Kenski et al., 2019, 2020; Schmid et al., 2022; Ziegele et al., 2020). The mentioned studies provide evidence that, for example, women, older people, and those scoring high on agreeableness typically experience antinormative utterances as more severe than men, younger people, and less agreeable persons do. Research also suggests that an increased use of social media promotes greater tolerance of incivil/intolerant comments due to distinct discussion norms as well as the prevalence and thus "normality" of antinormative utterances on these platforms (Haslop et al., 2021; Schmid et al., 2022; Ziegele et al., 2020).

To ensure the robustness of our results, we will thus conduct an exploratory multiverse analysis that considers the above-mentioned context-specific and person-specific factors in addition to the manipulated incivility and intolerance subtypes.

Method

We conducted a preregistered online factorial survey experiment with n =964 German participants in October 2022. Data, materials, and analysis scripts are available at https://doi.org/10.17605/OSF.IO/W92VJ; the preregistration is available at https://doi.org/10.17605/OSF.IO/EB9R3.

Design and procedure

Factorial survey experiment

The participants were confronted with four manipulated posts that were allegedly posted by users in response to recent news articles on an unspecified social media platform (see Figure 1). They were then asked to rate the posts' offensiveness and harm to society, and to indicate their intention to delete the post. The posts were created by systematically combining individual statements that either feature or do not feature characteristics of incivil (profanity | attacks toward arguments) and intolerant discourse (offensive stereotyping | violent threats) into coherent posts. In factorial survey terminology, posts can be understood as vignettes with four dimensions and two levels each, leading to a vignette universe of 16 vignettes (2 × 2 × 2 × 2 design).

Four orthogonal, balanced, and thus D-efficient sets of four vignettes each were created following literature recommendations (Dülmer, 2016). The participants were asked to rate only one randomly assigned vignette set, with vignettes displayed in a randomized order. After each post, participants were asked about their perceptions of the post and their intention to delete the post. Moreover, they were asked to indicate how similar they are to the creator of the post and to the group attacked/mentioned in the post. To prevent learning and fatigue as well as improve generalizability by accounting for possible influences of the topic of the post, we created posts for four different topics (see the next section). Topics were randomly assigned per participant, with each participant seeing a post on each topic. However, topics are not treated as an additional experimental factor, but are instead controlled for in the data analysis (Judd et al., 2012).

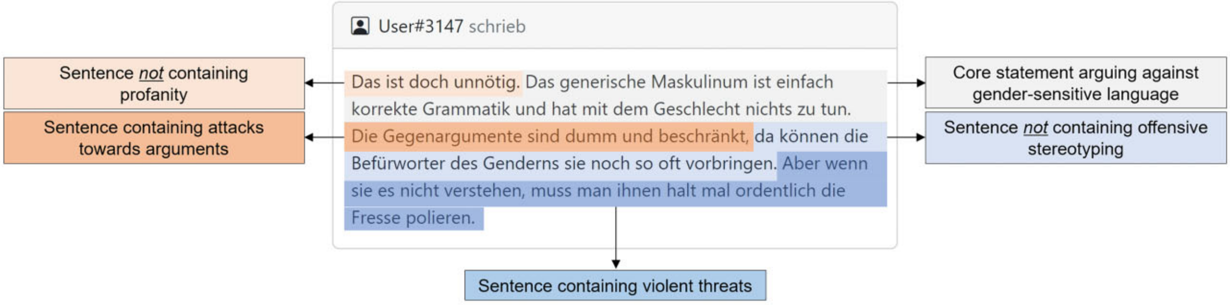
Stimuli

The four topics—(1) gender-sensitive language, (2) abortion, (3) migration, and (4) switching to renewable energies—were chosen for several reasons: Topics (1) and (4) are currently the subject of much discussion in the German media (Grimberg, 2022; von Pokrzywnicki, 2022), while Topics (2) and (3) can be characterized more as "long-running issues"[2] on which people in Germany tend to have rather stable opinions. Moreover, attitudes toward all four topics are likely to correlate with people's political attitudes as measured on a left–right spectrum. In order to provide a broad identification potential with the alleged posters in the stimulus, they hold a politically left-leaning opinion in the posts on Topics (2) and (4), while they hold a politically right-leaning opinion in the posts on Topics (1) and (3).

Each post consisted of a core statement that remained the same for every post on a particular topic, and four additional sentences/clauses that contained or did not contain profanity, attacks toward arguments, offensive stereotyping, and violent threats, respectively (see Figure 1). To obtain common arguments and wordings, we first examined online discussions about these four topics and consulted specialized websites that cover arguments to debatable issues (e.g., https://www.procon.org/). Moreover, we scanned the discussions for recurring forms of profanity, offensive stereotyping, etc. Building on the definitions of incivility and intolerance, statements that did or did not feature the four norm violations were then created for each topic. Particular emphasis was put on creating structurally similar statements of comparable length. All stimulus wordings and English translations can be obtained from the *Open Science Framework* (OSF) repository.

Qualitative expert pretest

Since perceptions of the posts are individual and depend both on personal characteristics and the specific configuration of incivility/intolerance, a standard pretest asking participants if


User#3147 schrieb
Sentence *not* containing profanity
Sentence containing attacks towards arguments
Core statement arguing against gender-sensitive language
Sentence *not* containing offensive stereotyping
Sentence containing violent threats

Das ist doch unnötig. Das generische Maskulinum ist einfach korrekte Grammatik und hat mit dem Geschlecht nichts zu tun. Die Gegenargumente sind dumm und beschränkt, da können die Befürworter des Genderns sie noch so oft vorbringen. Aber wenn sie es nicht verstehen, muss man ihnen halt mal ordentlich die Fresse polieren.

Figure 1. Example for a post used in the study.

Note: All used stimulus wordings and their translations are available in the OSF repository.

the post is incivil/intolerant is not suitable for our study. Instead, to check the adequacy of our operationalizations, we conducted qualitative interviews with five German-speaking researchers that are experts in the domain of incivility and not associated with this research project. They were confronted with our stimulus material and asked whether the operationalization was valid, whether the different combinations of the statements were plausible, and whether they had any other comments. Based on these interviews, we then revised the stimulus material for the study: (a) For the incivility subtype "attacks toward arguments," statements were adapted to ensure that attacks always clearly refer to arguments (and not to persons). (b) For the intolerance subtype "violent threats," a consistent definition of violence was used, so that in each case it is nonlethal, physical violence aimed at specific groups of people. (c) Last, the language was adjusted because it was partially perceived as "too academic."

Measures

Dependent variables

Offensiveness. After each post, participants were asked whether they agreed with three statements focusing on their perceived offensiveness ("The post is …" "offensive;" "hostile;" "hurtful") on a 7-point scale ranging from *does not apply at all* (1) to *does fully apply* (7), building on Saleem et al. (personal communication). Three additional statements focusing on the post's perceived adequacy ("…is accurate;" "…had to be said;" and "…is necessary") served as distractors. A mean index of the three items focusing on perceived offensiveness was then computed per participant and post ($M = 4.44$, $SD = 2.10$, $\omega_h = 0.93$).

Harm to society. Additionally, participants were asked whether they agreed with three statements focusing on the post's harm to society ("Posts like this…" "are harmful to society;" "threaten relationships between social groups;" "prevent a dialogue between social groups") on a 7-point scale, building on Saleem et al. (personal communication) and Leets (2001). Again, a mean index was computed per participant and post ($M = 4.22$, $SD = 2.02$, $\omega_h = 0.95$).

Deletion intention. Focusing more on behavioral aspects and forcing a clear decision, participants were also asked: "Suppose you had the authority to remove posts from the discussion. Would you delete this post?" They could then indicate whether they would do so ("yes") or not ("no"). Overall, the participants chose "yes" in 42.2% of all post evaluations.

Control variables

Perceived similarity to the creator of the post. After each post, the participants were also asked to assess how similar they perceived themselves to be to the creator of the post by responding to the statement: "Based on this post, how similar are you to the person who wrote the post?" on a 7-point scale ranging from *not at all similar* (1) to *very similar* (7), $M = 3.16$, $SD = 2.13$.

Perceived similarity to the group attacked in the post. Similarly, after each post, the participants were asked to assess how similar they perceived themselves to be to the group attacked/mentioned in the post (e.g., people supportive of gender-sensitive language in the posts arguing against gender-sensitive language) by responding to the statement: "And how similar are you to the group that is criticized in the post?" (same scale as above, $M = 2.75$, $SD = 1.96$).

Attitudes toward the topic of the post. Prior to exposure to the stimuli, to assess their own attitudes toward the four topics, the participants were asked to respond to one statement per topic (e.g., "Gender-sensitive language should be used in the media and in public communication") on a scale from *strongly disagree* (1) to *strongly agree* (7). Four additional statements on other topics served as distractors, with statements displayed in a randomized order per participant. Overall, the participants tended to be critical of gender-sensitive language ($M = 2.82$, $SD = 2.12$) and supportive of abortion rights ($M = 6.08$, $SD = 1.60$), while being somewhat more torn on the overall positive effects of migration ($M = 3.38$, $SD = 1.86$) and the switch to renewable energies ($M = 5.19$, $SD = 1.91$).

Satisfaction with the political system. The 4-item, 7-point German Satisfaction with the Political System Short Scale (SPS; Dentler et al., 2020) was used to measure the participants' political satisfaction, with higher values indicating more satisfaction (M=3.84, SD=1.54, ω_h=0.84).

Left–right self-placement. The participants' political attitudes were measured by using the one-item left–right self-placement scale (Breyer, 2015). Participants were asked to place themselves on a 9-point rating scale with the extreme poles labeled *left* (1) and *right* (9), $M = 4.89$, $SD = 1.69$.

Support for free speech. All further measures were obtained *after* exposure to all posts. Attitudes toward free speech were measured using the 3-item, 7-point scale developed by Riedl et al. (2021), with higher values indicating higher support for free speech ($M = 5.90$, $SD = 1.23$, $\omega_h = 0.89$).

Personality traits. The participants' personality traits were measured by using the Big Five Inventory-SOEP (BFI-S;

Schupp & Gerlitz, 2008). The scale measures neuroticism, extraversion, openness, conscientiousness, and compatibility with 15 items (three items for each personality trait) on a 7-point scale. Internal consistency was acceptable for extraversion ($M = 4.54$, $SD = 1.41$, $\omega_h = 0.77$) and neuroticism ($M = 4.02$, $SD = 1.41$, $\omega_h = 0.73$), but constricted for openness ($M = 4.79$, $SD = 1.29$, $\omega_h = 0.67$), conscientiousness ($M = 5.72$, $SD = 1.05$, $\omega_h = 0.66$), and agreeableness ($M = 5.39$, $SD = 1.06$, $\omega_h = 0.54$).

Sociodemographic variables and social media use. Finally, the participants reported various sociodemographic characteristics (see below) and their use of Facebook ($M = 5.01$, $SD = 2.98$), Twitter ($M = 2.40$, $SD = 2.33$), Instagram ($M = 4.26$, $SD = 3.15$), and TikTok ($M = 3.04$, $SD = 2.83$) on a 9-point scale ranging from *never* (1) to *almost every hour* (9) in a screening questionnaire placed at the beginning of the questionnaire. 14.9% never use any of those four social media platforms.

Participants

The participants were recruited through a commercial online access panel hosted by *Bilendi*. The sample was assembled using nationally representative quotas for age, gender, and education. All participants received financial compensation through the panel provider.

In total, 1,099 participants completed the questionnaire and passed an attention check placed before the post evaluations. Seven participants were excluded due to missing data; further 128 participants were excluded due to speeding or spending less than 10 s on the initial post evaluation page. The final sample of 964 participants was made up of 52.2% female, 45.7% male, and 0.1% non-binary persons, with age ranging from 18 to 86 (M=45.5 years, SD=15.1). 22.8% of participants hold a university degree.

Data preparation and statistical modeling

Several of the measures outlined above were modified for the data analysis. First, the multi-categorical variables gender and education were reduced to binary variables. Second, all metric predictors were mean-centered. Third, attitudes on the topic of the post were inverted for two attitude measures (gender-sensitive language, migration); thus, higher values for all attitude measures indicate higher agreement with the content of the post.

Because the participants evaluated four posts each, the dependent measures are not independent. Furthermore, stimulus wordings varied across topics. For all models, we thus estimate crossed random intercepts for both participants and topics. While we keep those models deliberately simple by only including binary predictors for profanity, attacks toward arguments, offensive stereotyping, and violent threats, we further investigate the robustness of the effects with a multiverse analysis (also known as specification curve analysis; Simonsohn et al., 2020; Steegen et al., 2016). A multiverse analysis seeks to identify the influence of conceptual and analytical decisions on the measured effects by identifying and modeling "the set of theoretically justified, statistically valid and non-redundant specifications" (Simonsohn et al., 2020, p. 1208). Such conceptual and analytical decisions may encompass the use of different operationalizations of predictors and outcomes, different subsets of samples, the (non-)inclusion of covariates, and the model estimation technique. All relevant specifications (i.e., reasonable combinations of these decisions) are modeled individually; the results then show the variance in the effects of interest (e.g., contrasts between factor levels) between different model specifications and provide insights into how particular specifications contribute to this variance.

Results

Confirmatory analyses: the effects of incivility and intolerance on perceptions of offensiveness, harm to society, and deletion intentions

We used linear and logistic multi-level regression models to test Hypotheses 1–5. The models included binary predictors for profanity, attacks toward arguments, offensive stereotyping, and violent threats. The results, shown in Table 1, indicate that there are no significant effects of profanity or attacks toward arguments on perceived offensiveness, perceived harm to society, and the intention to delete a post. However, there are significant positive effects of offensive stereotyping and violent threats on all three outcomes (i.e., the presence of these norm violations leads to an increase in perceived offensiveness, harm to society, and likelihood to report a deletion intention). To illustrate this, the third model predicts the chance of a participant to express the intention to delete a post at 17% [95% confidence interval (CI): 12%–22%] for posts containing only profanity, likewise at 17% (12%–23%) for posts containing only attacks on arguments, at 26% (19%–34%) for posts containing only offensive stereotyping, and at 52% (42%–61%) for posts containing only violent threats.[3] These results thus oppose *H1a–c* and *H2a–c*, which predicted an increased perceived offensiveness, perceived harm to society, and intention to delete for the presence of profanity and attacks toward arguments (i.e., incivil comments), and confirm *H3a–c* and *H4a–c*, which predicted the same effects for offensive stereotyping and violent threats (i.e., intolerant comments). Consequently, *H5*, which predicted stronger effects on the three outcomes for the intolerant norm violations than the incivil ones can be confirmed as well.

Exploratory analyses: multiverse analysis of conceptual and analytical specifications

We identified the following conceptual and analytical decisions that may affect the confirmatory results: First, we focused on three different outcomes [(1) perceived offensiveness, (2) perceived harm to society, and (3) deletion intention]. Second, the confirmatory analyses included only main effects of the four investigated norm violations, while the literature suggests that some norm violations may also work in tandem. We thus consider (1) the main effect-only model as well as (2) a model that also includes all possible two-way interactions between the four norm violations. Third, we account for topic effects by either (1) including random intercepts per topic or (2) also estimating random slopes for all four norm violations across topics. Fourth, we fit these models based on (1) the final sample as outlined above, (2) excluding all non-users of social media in this sample, and (3) expanding the final sample with participants excluded for speeding to account for differences in perceptions related to social media use habits (see above) and the inherent arbitrariness of speeding cutoffs. Last, we examine the influence of control variables on the effects of the four norm violations on the outcomes by (1) including no control variables (as in the

Table 1. Regression models predicting offensiveness, harm to society, and deletion intention

Model	(1) Offensiveness		(2) Harm to society		(3) Deletion intention	
Fixed effects	Coef.	95% CI	Coef.	95% CI	Coef.	95% CI
Incivility						
Profanity	0.11	[−0.01 to 0.22]	0.03	[−0.08 to 0.15]	0.04	[−0.13 to 0.21]
Attacks toward arguments	0.10	[−0.01 to 0.20]	0.03	[−0.08 to 0.13]	0.04	[−0.12 to 0.20]
Intolerance						
Offensive stereotyping	0.77	[0.66–0.88]	0.43	[0.32–0.53]	0.59	[0.43–0.76]
Violent threats	1.45	[1.35–1.56]	1.14	[1.04–1.24]	1.73	[1.55–1.90]
Variance components						
σ (Participant)	1.01		1.01		1.24	
σ (Topic)	0.35		0.39		0.33	
σ (Residuals)	1.63		1.61			
Model fit						
Conditional R^2	0.41		0.37		0.43	
Marginal R^2	0.15		0.09		0.14	
AIC	15,646.26		15,576.14		4609.45	

Notes: Linear (Models 1 and 2) and logistic (Model 3) multilevel regression models with random intercepts for participants and topics. $n_{\text{Posts}} = 3{,}856$, $n_{\text{Participants}} = 964$, $n_{\text{Topics}} = 4$. Coefficients are unstandardized regression coefficients. R^2 for generalized linear multilevel models as proposed by Nakagawa et al. (2017).

confirmatory analyses), (2–11) including 10 different, potentially relevant context- or person-specific covariates (see the section *The influence of context- and person-specific factors*) one at a time, and (12) including all 10 covariates at the same time in the models. This leads to 3 × 2 × 2 × 3 × 12 = 432 model specifications estimated in the multiverse analysis.

The results of the 288 model specifications predicting perceptions of offensiveness and harm to society are displayed in Figures 2a and b and 3a and b.[4] In general, the results support the assumption that intolerant norm violations lead to stronger effects than incivil norm violations. Across all specifications, the effects of offensive stereotyping and violent threats (i.e., the intolerant norm violations) are positive (i.e., leading to higher perceived offensiveness and harm to society) and significant in 100% of all cases (see Figure 3a). Contrariwise, the effects of profanity and attacks toward arguments (i.e., the incivil norm violations) are positive and significant in only 46.5% and 37.2% of call cases, respectively (see Figure 2a).

Regarding the effects of the individual specifications, the multiverse analysis revealed that when controlling for the perceived similarity to the person who violated social norms (the alleged poster), the effects of norm violations were generally smaller. This suggests that whether someone feels offended or perceives harm to society depends on whether they perceive themselves to belong to the same social group as the norm violator or not. We also find that all four norm violations affected perceived offensiveness more than perceived harm to society. This was especially the case for profanity (ratio of median effect size: 2.19) and attacks toward arguments (2.32), where contrasts between a post containing either norm violation and posts not containing either norm violation were, on average, more than twice as large for the perceived offensiveness measure when compared with the perceived harm to society measure.

Overall, the variance in the effect size for both intolerant norm violations was mostly due to the different outcomes (perceived offensiveness vs. perceived harm to society), which accounted for 89.5% (offensive stereotyping) and 62.3% (violent threats) of the total effect size variance. In contrast, the relationship between the predictors (main effect-only models vs. two-way interaction models) explained the majority of effect size variance for profanity (71.7%) and to a lesser degree for attacks toward arguments (41.1%).[5] This suggests that these two incivil norm violations mostly explain variance in our measured outcomes if their interactions with the intolerant norm violations are accounted for.

Finally, as all 10 investigated control variables were included in 72 models each, we can also share some exploratory findings on the effects of these control variables on the three measured outcomes. Higher perceived similarity to the alleged poster consistently predicted a *decrease* in perceived offensiveness, perceived harm to society, and deletion intention (100% of all effects negative and significant). Likewise, higher perceived similarity to the group attacked in the post consistently predicted an *increase* in the three outcome measures (100% of all effects positive and significant). The effects of the other control variables were less consistent, with only users' attitude on the topic of the post and their satisfaction with the political system having the same effect in more than 80% of all models: A more similar attitude to the stance expressed in the post predicted a significantly lower perception of offense, harm to society, or deletion intention in 83% of all models the covariate was included in, whereas a higher satisfaction with the political system predicted significant increases in the outcome measures in 93% of all models. Less consistent effect estimates were found for the left–right self-placement, support for free speech, agreeableness, as well as age, gender, and formal education.[6]

Discussion

Building on recent research that challenges the notion that incivility in online discussions is inherently detrimental (Oh et al., 2021; Rossini, 2019, 2022), this study relies on a distinction between *incivil* and *intolerant* user comments and investigates how online users perceive and react to these distinct forms of antinormative discourse online. To do so, we have conducted a preregistered factorial survey experiment with a nationally representative sample of n = 964 German online users and presented them with manipulated user comments that included statements associated with incivil (profanity; attacks toward arguments) and intolerant discourse (offensive stereotyping; violent threats).

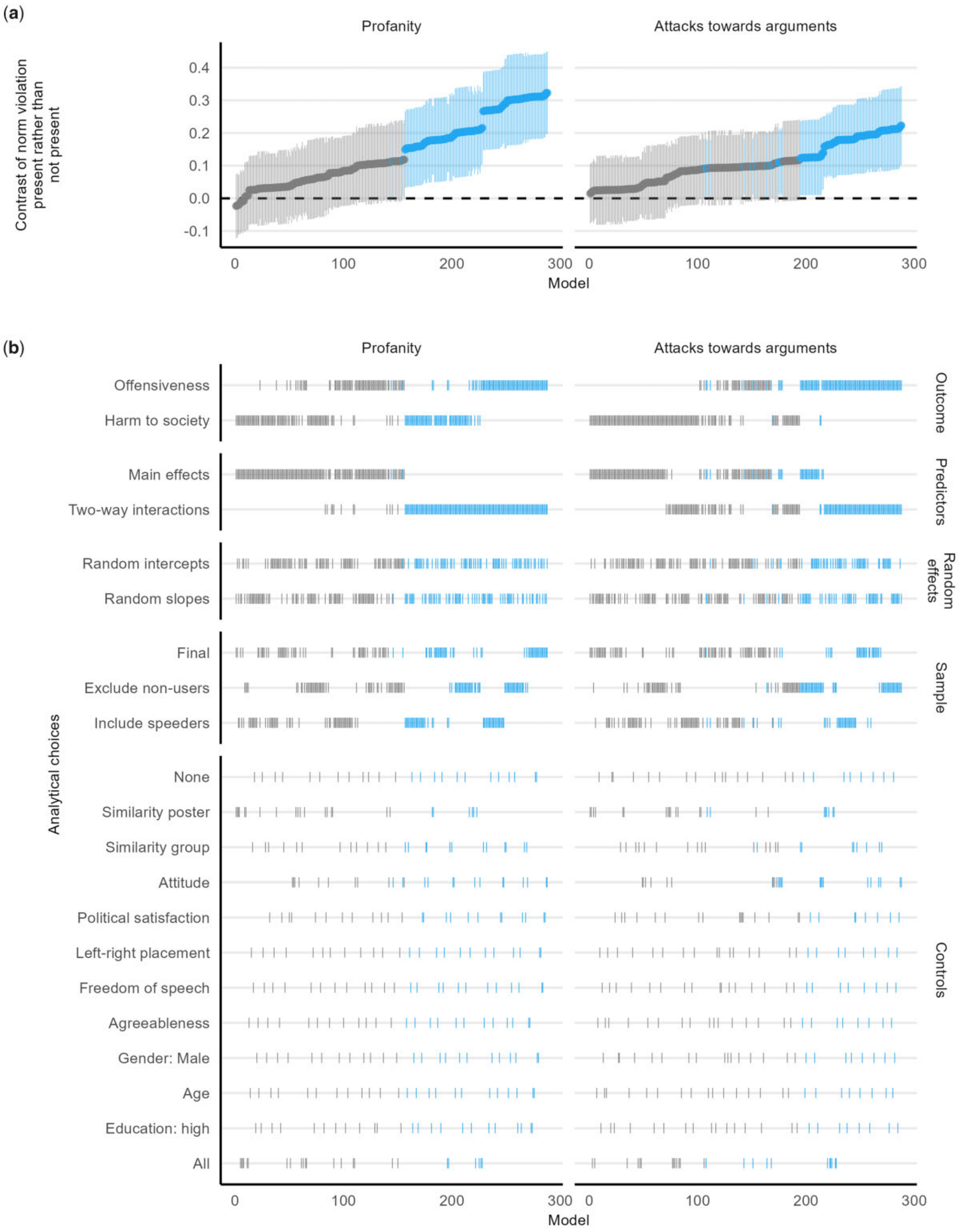


Figure 2. Multiverse analysis of the effects of incivility (288 specifications).

Notes: The upper panel (a) shows the specification curve as contrasts and their 95% confidence intervals, estimated at sample means and the response scale, of norm violations present rather than not present. The lower panel (b) shows the effects of the individual analytical choices on the contrasts, with each tick representing one model. Negative significant contrasts [i.e., 95% confidence intervals (CIs) not overlapping 0] are plotted in red, non-significant contrasts in gray, and positive significant contrasts in blue.

The findings provide evidence that—in line with content-analytical research and the theoretical distinction between incivility as a primarily rhetorical asset and intolerance as a behavior that threatens democratic values (Oh et al., 2021; Rossini, 2019, 2022)—participants consistently perceive intolerant statements as more offensive and more harmful to society than incivil ones. Likewise, they report a higher intention to delete comments that contain intolerance when compared with incivility. An exploratory multiverse analysis demonstrates that the effects are robust across a range of

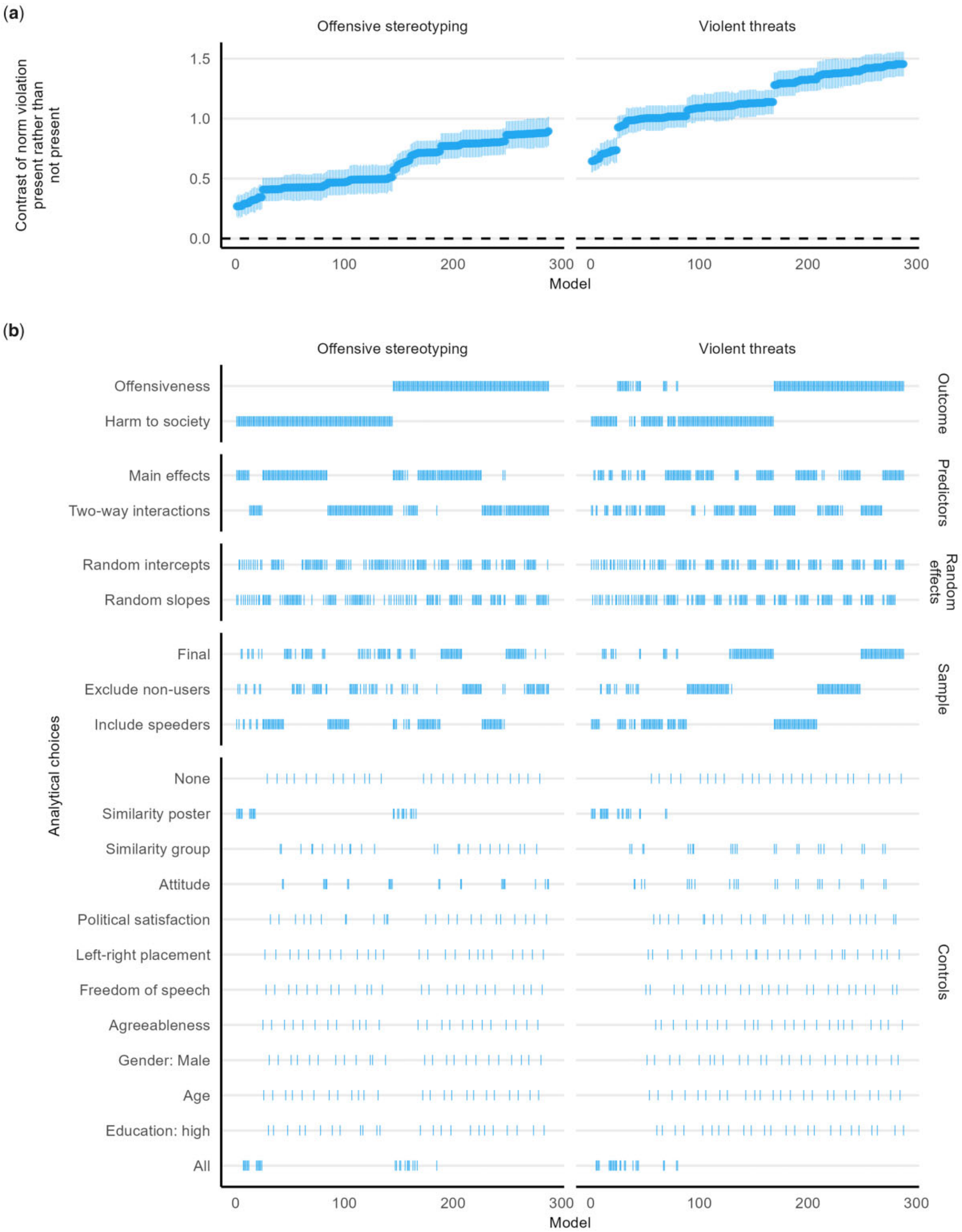


Figure 3. Multiverse analysis of the effects of intolerance (288 specifications).

Notes: The upper panel (a) shows the specification curve as contrasts and their 95% CIs, estimated at sample means and the response scale, of norm violations present rather than not present. The lower panel (b) shows the effects of the individual analytical choices on the contrasts, with each tick representing one model. Negative significant contrasts (i.e., 95% CIs not overlapping 0) are plotted in red, non-significant contrasts in gray, and positive significant contrasts in blue.

analytical decisions, including different model specifications and adjusting for the influence of theoretically relevant context- and person-specific factors. Thus, our results also reflect previous studies that have shown that intolerant behaviors—encouraging harm and using racial or sexual slurs—are perceived markedly worse than incivil ones such as resorting to vulgarity or attacking one's stand on issues (Stryker et al., 2016, 2022). The strong effects of intolerant utterances on participants' deletion intentions furthermore support research showing that hateful comments are more likely to get flagged than "mere" disparaging comments (Kunst et al., 2021) and that calls for violence in comments drastically increase users' reporting intentions (Wilhelm et al., 2020). Our findings are important for understanding democratic discourse online, as they suggest that intolerant comments can severely impair the quality of online discussions. Such comments create an environment that is perceived as hostile to individuals and social groups, which in turn hinders the ability to engage in respectful and constructive debates, especially on controversial political issues. The observed effects on deletion intentions furthermore suggest that users may support moderation policies aimed at curbing threatening or harmful speech, highlighting the need for platform providers to address these types of norm violations in a targeted manner.

The multiverse analysis revealed that the specific dependent variable measured is the central source of variation in the effects of the four incivil/intolerant norm violations. This means that the impact of incivil and intolerant comments varies depending on the outcome of interest. Whether people are considering their personal outrage, danger to others, or the comment's relevance to the discussion, the effects differ. For instance, we consistently found stronger effects on perceived offensiveness than on perceived harm to society, suggesting that personal considerations are more affected than societal ones. The different outcomes were by far the main source of variation for the effects of the intolerant norm violations, which suggests that other modeling decisions were less relevant here. Conversely, the effects of the incivil norm violations were more dependent on other modeling decisions, especially the relationship between the predictors: Both profanity and attacks toward arguments (i.e., the incivil norm violations) were more likely to affect the outcomes when their interactions with intolerance were considered. Thus, incivil norm violations may act more as an "intensifier," whereas intolerant norm violations seem to affect readers' perceptions and reactions more globally.

From a methodological perspective, our findings also have implications for computational approaches of incivility detection. Large-scale, automatic investigations of online incivility often rely on binary classifiers (Theocharis et al., 2020; Unkel & Kümpel, 2022) that subsume several (sub-)dimensions of incivility and intolerance under one umbrella. While these investigations may adequately capture trends in discussion *tone*, they are not able to distinguish between the conceptually and empirically different notions and nuances in discussion *substance* that incivil and intolerant norm violations entail. At the moment, it seems easier to automatically detect incivil rather than intolerant utterances, as especially the subtype of profanity is often rooted in single (e.g., four-letter) words, whereas intolerant norm violations such as offensive stereotyping are usually more subtle and contextual (e.g., Stoll et al., 2020). This is not only relevant for computational content-analytical research, but also for practitioners (e.g., community managers), because automated moderating tools are more likely to detect incivil rather than intolerant comments—though the latter are by far the more serious problem and, as our study shows, also perceived as such by users.

The generalizability of our results is subject to certain limitations. First, while the factorial survey design guaranteed high internal validity and enabled us to systematically investigate the influence of the four subtypes of incivility/intolerance, the external validity is compromised. For methodological reasons, each norm violation (e.g., profanity) was located in an individual statement, albeit content analyses have shown that incivility and intolerance are not independent of each other in real online discussions (Oh et al., 2021; Rossini, 2022). Specifically, these studies suggest that intolerant comments are more likely to be both intolerant *and* incivil (e.g., because violent threats are coupled with profanity). While on the level of the comment, co-occurrences of incivility and intolerance were possible in our study, individual statements had to be designed to only reflect one specific subtype. Moreover, although the literature identifies more subtypes of incivility/intolerance than we examined, we had to focus on two subtypes each due to resource constraints and avoiding an "overload" of the stimulus, as each additional subtype would have required an additional clause in the post. Future research should extend our findings by testing whether other subtypes of incivility/intolerance show comparable effects on users' perceptions and reactions.

Second, we measured our outcome variables only for single comments, thus ignoring possible effects on the *dynamics* of online discussions. While, for example, profanity was perceived as "less bad" in our study, it might serve as a gateway for worse norm violations in the further course of the discussion. Indeed, previous research has shown that "incivility foments incivility" (Chen & Lu, 2017, p. 121), with norm-violating comments prompting more norm-violating replies in CMC settings (see also Kim et al., 2021; Shmargad et al., 2022; Unkel & Kümpel, 2022). Future research could validate our findings by integrating the manipulated comments into a functional comment section that enables actual user engagement (for such an approach see, e.g., Kalch & Naab, 2018).

Third, our participants were exposed to posts from an undefined social media platform, with the only deductible information being that the platform affords a high degree of *anonymity*—posters were only identified by names like "User#1234" and nondescript icons. However, research suggests that differences in perceived affordances between social media platforms are related to varying perceptions of the prevalence of incivility (Sude & Dvir-Gvirsman, 2023), implying that users' perceptions of offensiveness or harm to society may be affected as well. Especially when studying existing social media platforms or communities, researchers need to consider how their unique architectures and platform norms influence how incivil/intolerant utterances are evaluated (see also Rieger et al., 2021).

Fourth, we conducted our study in Germany, which is why our results may not be generalizable to other countries and/or cultural contexts. Indeed, Germany has one of the most restrictive anti-hate speech laws (Hawdon et al., 2017), which could explain the observed strong effects of the intolerant norm violations in particular. While first cross-national experimental research—comparing the Netherlands, UK, and Spain—suggests that "some, if not most, forms of incivility are *not* dependent on country-context but rather unacceptable everywhere" (Otto et al., 2020, p. 101), results might look different in other countries. For instance, in the United States, the significance of freedom of speech may lead to a more lenient approach to verbal norm violations and thus less negative perceptions. Accordingly, we encourage other researchers to conduct replications of our study in different cultural contexts.

Notwithstanding these limitations, our study provides important insights into how online users perceive and react to distinct forms of antinormative discourse. By adding the perspective of the users, the findings support the assumption that it is not incivility, but intolerance that threatens the quality and substance of political discussions on social media platforms.

Data availability

The data underlying this article are available on the *Open Science Framework* (OSF), at https://doi.org/10.17605/OSF.IO/W92VJ.

Conflicts of interest: None declared.

Open science framework badges

Open Materials

The components of the research methodology needed to reproduce the reported procedure and analysis are publicly available for this article.

Open Data

Digitally shareable data necessary to reproduce the reported results are publicly available for this article.

Preregistered

Research design was preregistered.

Notes

1. Prior research has repeatedly shown the high prevalence of vulgarity and *profanity* in online comments (Coe et al., 2014; Oh et al., 2021; Sood et al., 2012), which is often used to lend weight to an argument or when "ranting" on a topic. While seemingly not the most prevalent type of incivility (Rossini, 2019), *attacks toward arguments* were particularly pertinent to our thematic focus on political discussions associated with distinct pro and con arguments. Compared with valid criticism, such incivil attacks do not simply question the appropriateness of an argument, but aim at invalidating and verbally discrediting it.
2. Abortion is and remains a hot topic in many countries of the EU (Grimm & Stavenhagen, 2016) and has received renewed attention in 2022, because the German Bundestag ended the ban on the advertisement of abortion services. Migration is routinely named one of the "most important problems" in the long-standing German election poll *Politbarometer* and was also among the four topics with the highest media presence in 2022 (von Pokrzywnicki, 2022).
3. For comparison: The chance to express the intention to delete is at 16% (11%–22%) for posts containing no incivil and intolerant expressions, and at 67% (59%–76%) for posts containing all four norm violations.
4. As our third outcome deletion intentions necessitates logistic instead of linear models, coefficients and effects are not directly comparable with the models on perceived offensiveness/harm to society. The analysis of the remaining 144 specifications focusing on deletion intentions can be found in the OSF repository, with results being similar to the ones outlined for the two perceptual outcomes.
5. Again, this pattern also holds for deletion intentions.
6. See Figure A3 in the OSF repository file OSF_Appendix.pdf for an overview of all control variable effect estimates across all model specifications.

Acknowledgments

The authors would like to thank Lena Frischlich, Magdalena Obermaier, Sünje Paasch-Colberg, Svenja Schäfer, and Marc Ziegele for taking part in the qualitative expert pretest. Their valuable feedback was instrumental in refining the stimuli for our study. Moreover, they would like to thank the Editor-in-Chief and the three anonymous reviewers for their time and effort in evaluating our manuscript. Their insightful comments and constructive feedback have greatly improved the quality of this article.

References

Borah, P. (2014). Does it matter where you read the news story? Interaction of incivility and news frames in the political blogosphere. *Communication Research*, *41*(6), 809–827. https://doi.org/10.1177/0093650212449353

Bormann, M. (2022). Perceptions and evaluations of incivility in public online discussions—Insights from focus groups with different online actors. *Frontiers in Political Science*, *4*, 812145. https://doi.org/10.3389/fpos.2022.812145

Bormann, M., Tranow, U., Vowe, G., & Ziegele, M. (2022). Incivility as a violation of communication norms—A typology based on normative expectations toward political communication. *Communication Theory*, *32*(3), 332–362. https://doi.org/10.1093/ct/qtab018

Breyer, B. (2015). Left–right self-placement (ALLBUS). *ZIS—The Collection of Items and Scales for the Social Sciences*. https://doi.org/10.6102/ZIS83

Brown, A. (2018). What is so special about online (as compared to offline) hate speech? *Ethnicities*, *18*(3), 297–326. https://doi.org/10.1177/1468796817709846

Chen, G. M., & Lu, S. (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. https://doi.org/10.1080/08838151.2016.1273922

Chen, G. M., Muddiman, A., Wilner, T., Pariser, E., & Stroud, N. J. (2019). We should not get rid of incivility online. *Social Media + Society*, *5*(3), 1–5. https://doi.org/10.1177/2056305119862641

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679. https://doi.org/10.1111/jcom.12104

Costello, M., Hawdon, J., Bernatzky, C., & Mendes, K. (2019). Social group identity and perceptions of online hate. *Sociological Inquiry*, *89*(3), 427–452. https://doi.org/10.1111/soin.12274

Dentler, K., Bluemke, M., & Gabriel, O. W. (2020). German satisfaction with the political System Short Scale (SPS). *ZIS—The Collection of Items and Scales for the Social Sciences*. https://doi.org/10.6102/ZIS278

Dülmer, H. (2016). The factorial survey: Design selection and its impact on reliability and internal validity. *Sociological Methods & Research*, *45*(2), 304–347. https://doi.org/10.1177/0049124115582269

Fraser, B. (1990). Perspectives on politeness. *Journal of Pragmatics*, *14*(2), 219–236. https://doi.org/10.1016/0378-2166(90)90081-N

Gagrčin, E. (2022). Your social ties, your personal public sphere, your responsibility: How users construe a sense of personal responsibility for intervention against uncivil comments on Facebook. *New Media & Society*. Advance Online Publication. https://doi.org/10.1177/14614448221117499

Grimberg, S. (2022, January 14). Gendergerechte Sprache—Darum geht's! *MDR Medien 360G*. https://www.mdr.de/medien360g/medienwissen/gendergerechte-sprache-110.html

Grimm, R., & Stavenhagen, L. (2016). *Einstellungen und Meinungen zum Schwangerschaftsabbruch in Europa—eine vergleichende Studie*. Ipsos Deutschland. https://www.ipsos.com/sites/default/files/2017-03/WP_Schwangerschaftsabbruch%20in%20Europa_RZ.pdf

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

Haslop, C., O'Rourke, F., & Southern, R. (2021). #NoSnowflakes: The toleration of harassment and an emergent gender-related digital divide, in a UK student online culture. *Convergence*, *27*(5), 1418–1438. https://doi.org/10.1177/1354856521989270

Hawdon, J., Oksanen, A., & Räsänen, P. (2017). Exposure to online hate in four nations: A cross-national consideration. *Deviant*

Behavior, *38*(3), 254–266. https://doi.org/10.1080/01639625.2016.1196985

Herbst, S. (2010). *Rude democracy: Civility and incivility in American politics*. Temple University Press.

Jay, T., & Janschewitz, K. (2008). The pragmatics of swearing. *Journal of Politeness Research*, *4*(2), 267–288. https://doi.org/10.1515/JPLR.2008.013

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Kalch, A., & Naab, T. K. (2018). Replying, disliking, flagging: How users engage with uncivil and impolite comments on news sites. *SCM Studies in Communication and Media*, *6*(4), 395–419. https://doi.org/10.5771/2192-4007-2017-4-395

Kenski, K., Coe, K., & Rains, S. A. (2019). Perceptions of incivility in public discourse. In R. G. Boatright, T. J. Shaffer, S. Sobieraj, & D. G. Young (Eds.), *A crisis of civility*? (pp. 45–60). Routledge.

Kenski, K., Coe, K., & Rains, S. A. (2020). Perceptions of uncivil discourse online: An examination of types and predictors. *Communication Research*, *47*(6), 795–814. https://doi.org/10.1177/0093650217699933

Kim, J. W., Guess, A., Nyhan, B., & Reifler, J. (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication*, *71*(6), 922–946. https://doi.org/10.1093/joc/jqab034

Kim, J. W., & Park, S. (2019). How perceptions of incivility and social endorsement in online comments (dis)encourage engagements. *Behaviour & Information Technology*, *38*(3), 217–229. https://doi.org/10.1080/0144929X.2018.1523464

Kim, Y. (2022). *Potentials and limitations of computer-mediated communication theories for online incivility research: A focus on bystander dynamics*. Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2022.761

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Kunst, M., Porten-Cheé, P., Emmer, M., & Eilders, C. (2021). Do "good citizens" fight hate speech online? Effects of solidarity citizenship norms on user responses to hate comments. *Journal of Information Technology & Politics*, *18*(3), 258–273. https://doi.org/10.1080/19331681.2020.1871149

Lee, J., Choi, J., & Kim, J. (2022). Effects of online incivility and emotions toward in-groups on cross-cutting attention and political participation. *Behaviour & Information Technology*, *41*(14), 3013–3027. https://doi.org/10.1080/0144929X.2021.1969429

Leets, L. (2001). Explaining perceptions of racist speech. *Communication Research*, *28*(5), 676–706. https://doi.org/10.1177/009365001028005005

Muddiman, A. (2017). Personal and public levels of political incivility. *International Journal of Communication*, *11*(2017), 3182–3202.

Muddiman, A. (2019). How people perceive political incivility. In R. G. Boatright, T. J. Shaffer, S. Sobieraj, & D. G. Young (Eds.), *A crisis of civility?* (pp. 31–44). Routledge. https://doi.org/10.4324/9781351051989-3/

Muddiman, A., & Stroud, N. J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication*, *67*(4), 586–609. https://doi.org/10.1111/jcom.12312

Mutz, D. C. (2015). *In-your-face politics: The consequences of uncivil media*. Princeton University Press.

Naab, T. K., Naab, T., & Brandmeier, J. (2021). Uncivil user comments increase users' intention to engage in corrective actions and their support for authoritative restrictive actions. *Journalism & Mass Communication Quarterly*, *98*(2), 566–588. https://doi.org/10.1177/1077699019886586

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*(134), 1–11. https://doi.org/10.1098/rsif.2017.0213

Oh, D., Elayan, S., Sykora, M., & Downey, J. (2021). Unpacking uncivil society: Incivility and intolerance in the 2018 Irish abortion referendum discussions on Twitter. *Nordicom Review*, *42*(1), 103–118. https://doi.org/10.2478/nor-2021-0009

Otto, L. P., Lecheler, S., & Schuck, A. R. T. (2020). Is context the key? The (non-)differential effects of mediated incivility in three European countries. *Political Communication*, *37*(1), 88–107. https://doi.org/10.1080/10584609.2019.1663324

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, *6*(2), 259–283. https://doi.org/10.1177/1461444804041444

Pennington, N., & Winfrey, K. L. (2021). Engaging in political talk on Facebook: Investigating the role of interpersonal goals and cognitive engagement. *Communication Studies*, *72*(1), 100–114. https://doi.org/10.1080/10510974.2020.1819844

Rega, R., & Marchetti, R. (2021). The strategic use of incivility in contemporary politics. The case of the 2018 Italian general election on Facebook. *The Communication Review*, *24*(2), 107–132. https://doi.org/10.1080/10714421.2021.1938464

Riedl, M. J., Naab, T. K., Masullo, G. M., Jost, P., & Ziegele, M. (2021). Who is responsible for interventions against problematic comments? Comparing user attitudes in Germany and the United States. *Policy & Internet*, *13*(3), 433–451. https://doi.org/10.1002/poi3.257

Rieger, D., Kümpel, A. S., Wich, M., Kiening, T., & Groh, G. (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), 1–14. https://doi.org/10.1177/20563051211052906

Rossini, P. (2019). Disentangling uncivil and intolerant discourse in online political talk. In R. G. Boatright, T. J. Shaffer, S. Sobieraj, & D. G. Young (Eds.), *A crisis of civility?* (pp. 142–157). Routledge. https://doi.org/10.4324/9781351051989-9

Rossini, P. (2022). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, *49*(3), 399–425. https://doi.org/10.1177/0093650220921314

Schmid, U. K., Kümpel, A. S., & Rieger, D. (2022). How social media users perceive different forms of online hate speech: A qualitative multi-method study. *New Media & Society*. Advance Online Publication. https://doi.org/10.1177/14614448221091185

Schupp, J., & Gerlitz, J.-Y. (2008). Big Five Inventory-SOEP (BFI-S). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. https://doi.org/10.6102/ZIS54

Shmargad, Y., Coe, K., Kenski, K., & Rains, S. A. (2022). Social norms and the dynamics of online incivility. *Social Science Computer Review*, *40*(3), 717–735. https://doi.org/10.1177/0894439320985527

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Sood, S., Antin, J., & Churchill, E. (2012). Profanity use in online communities. In *Proceedings of the SIGCHI conference on human factors in computing systems (CHI'12)*. ACM, New York, NY, USA, 1481–1490. https://doi.org/10.1145/2207676.2208610

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Stoll, A., Ziegele, M., & Quiring, O. (2020). Detecting impoliteness and incivility in online discussions: Classification approaches for German user comments. *Computational Communication Research*, *2*(1), 109–134. https://doi.org/10.5117/CCR2020.1.005.KATH

Stryker, R., Conway, B. A., Bauldry, S., & Kaul, V. (2022). Replication note: What is political incivility? *Human Communication Research*, *48*(1), 168–177. https://doi.org/10.1093/hcr/hqab017

Stryker, R., Conway, B. A., & Danielson, J. T. (2016). What is political incivility? *Communication Monographs*, *83*(4), 535–556. https://doi.org/10.1080/03637751.2016.1201207

Su, L. Y.-F., Xenos, M. A., Rose, K. M., Wirz, C., Scheufele, D. A., & Brossard, D. (2018). Uncivil and personal? Comparing patterns of incivility in comments on the Facebook pages of news outlets. *New Media & Society*, *20*(10), 3678–3699. https://doi.org/10.1177/1461444818757205

Sude, D. J., & Dvir-Gvirsman, S. (2023). Different platforms, different uses: Testing the effect of platforms and individual differences on perception of incivility and self-reported uncivil behavior. *Journal of Computer-Mediated Communication*, *28*(2), 1–13. https://doi.org/10.1093/jcmc/zmac035

Theocharis, Y., Barberá, P., Fazekas, Z., & Popa, S. A. (2020). The dynamics of political incivility on Twitter. *SAGE Open*, *10*(2), 1–15. https://doi.org/10.1177/2158244020919447

Unkel, J., & Kümpel, A. S. (2022). Patterns of incivility on U.S. congress members' social media accounts: A comprehensive analysis of the influence of platform, post, and person characteristics. *Frontiers in Political Science*, *4*, 1–11. https://doi.org/10.3389/fpos.2022.809805

von Pokrzywnicki, U. (2022). PMG Themenrennen—Diese Themen bewegen Deutschland. *PMG Presse-Monitor*. https://www.pressemonitor.de/blog/pmg-themenrennen-diese-themen-bewegen-deutschland/

Wilhelm, C., Joeckel, S., & Ziegler, I. (2020). Reporting hate comments: Investigating the effects of deviance characteristics, neutralization strategies, and users' moral orientation. *Communication Research*, *47*(6), 921–944. https://doi.org/10.1177/0093650219855330

Williams, A., Oliver, C., Aumer, K., & Meyers, C. (2016). Racial microaggressions and perceptions of Internet memes. *Computers in Human Behavior*, *63*, 424–432. https://doi.org/10.1016/j.chb.2016.05.067

Ziegele, M., Naab, T. K., & Jost, P. (2020). Lonely together? Identifying the determinants of collective corrective action against uncivil comments. *New Media & Society*, *22*(5), 731–751. https://doi.org/10.1177/1461444819870130

Ziegele, M., Springer, N., Jost, P., & Wright, S. (2018). Online user comments across news and other content formats: Multidisciplinary perspectives, new directions. *Studies in Communication and Media (SCM)*, *6*(4), 315–332. https://doi.org/10.5771/2192-4007-2017-4-315