Article

Communication Research
1–24

DOI: 10.1177/00936502211062773
journals.sagepub.com/home/crx


Is Constructive Engagement Online a Lost Cause? Toxic Outrage in Online User Comments Across Democratic Political Systems and Discussion Arenas

Julia Jakob[1], Timo Dobbrick[1], Rainer Freudenthaler[1], Patrik Haffner[2], and Hartmut Wessler[1]

Abstract

This study is the first to simultaneously investigate country-level and platform-related context factors of toxic outrage, that is, destructive incivility, in online discussions. It compares user comments on the public role of religion and secularism from 2015/16 in four democracies (Australia, United States, Germany, Switzerland) and four discussion arenas on three platforms (News websites, Facebook, Twitter). A novel automated content analysis ($N = 1{,}236{,}551$) combines LIWC dictionaries with machine learning. The level of toxic outrage is higher in majoritarian than in consensus-oriented democracies and in arenas that afford plural, issue-driven rather than like-minded, preference-driven debates. Yet, toxic outrage is lower in forums that tend to separate public and private conversations than in those that collapse varying contexts. This suggests that user-generated discussions flourish in environments that incentivize actors to strive for compromise, put relevant issues center stage and make room for public debate at a relative distance from purely social conversation.

Keywords

toxic outrage, incivility, online discussion, political system, socio-technical affordances

[1]University of Mannheim, Baden-Württemberg, Germany
[2]Forum Institut für Management GmbH, Heidelberg, Baden-Württemberg, Germany

Corresponding Author:
Julia Jakob, Mannheim Centre for European Social Research, University of Mannheim, A5, 6 – Building A, Mannheim, Baden-Württemberg 68159, Germany.
Email: julia.jakob@mzes.uni-mannheim.de

Digital spaces provide an infrastructure for abusive discourse to spread more quickly and broadly than ever (Coe et al., 2014). Up to 40% of the contributions in online discussions can contain uncivil elements (Ziegele et al., 2018). Civility as an expression of mutual respect has long played a central role in online public sphere research and is a central normative dimension to assess the democratic quality of online discussions (Friess & Eilders, 2015). This is rooted either in deliberative theory, which postulates that mutual respect increases the openness for opposing arguments (Kies, 2010), or in liberal notions of communicative restraint that aim to prevent social conflicts from escalating (Ackerman, 1989). Yet, uncivil language can also serve minority groups in public discussions who are otherwise not heard at all (Jamieson et al., 2017). Should normative conceptions of democratic public discourse thus allow incivility as a generally acceptable component of online speech in a liberal-individualist fashion (Freelon, 2015)?

The danger of this "anything goes" position is that normative analysis of online speech would lose its bite exactly at a time when democracy is increasingly imperiled. Even relaxed standards for online communication need to draw a line between constructive and less constructive debate contributions (Bächtiger et al., 2010). This study argues that what distinguishes normatively acceptable from unacceptable forms of incivility is that the latter generate a fundamental insensitivity toward other perspectives. It investigates toxic outrage as a rhetorical strategy that fosters such insensitivity by aiming at provoking negative emotional reactions in the audience, thus promoting closed-mindedness toward political opponents.

To limit its expansion, it is key to understand which context factors drive toxic outrage online and inhibit the more constructive styles of user-generated debate needed in healthy democracies. While studies increasingly investigate the impact of sociotechnical affordances (Nagy & Neff, 2015) on debate quality (e.g., Esau et al., 2017; Freelon, 2015; Rowe, 2015b), country-level structural influences are rarely explored (Ruiz et al., 2011). In considering national political and platform-related antecedents of toxic outrage in online debates together, this study is one of the first to examine the phenomenon in its multi-layered environment.

In an automated content analysis combining dictionary-based analysis with machine learning, we investigate how the *political system* of a country (Lijphart, 2012), the degree of *context collapse* in (Boyd, 2011) and the *primary use function* of a discussion arena (Maia & Rezende, 2016) condition toxic outrage online. Specifically, we study user posts on the public role of religion and secularism in society from August 2015 to July 2016 in two majoritarian and two consensus-oriented democracies, namely Australia, the United States, Germany and Switzerland. We compare user comments from four discussion arenas on three platforms, namely from (a) mainstream news media's website comment sections and (b) their Facebook pages, from the (c) Facebook pages of partisan collective actors and alternative media and from (d) Twitter.

Theory

Incivility Across Democratic Theories

Holmes (1988) advocates the strictest civility norm. From his liberal perspective, rules of omission can be constructive for democratic debates because "by tying our tongues

about a sensitive question, we can secure forms of cooperation and fellowship otherwise beyond reach" (p. 19). Even in softer forms, this conversational restraint norm fundamentally aims at preventing conflicts from escalating: If a topic is controversial, we should "simply say nothing at all about this disagreement and put the moral ideals that divide us off the conversational agenda" (Ackerman, 1989, p. 16) so that respectful political collaboration can continue (Rawls, 1987). Liberal theorists remind us that social situations themselves often incentivize self-restraint: Do we not often find ourselves in debates with acquaintances where we end up unable to constructively debate the topic and decide to drop it (Ackerman, 1989; Holmes, 1988)? This also used to be the case for traditional mass media. Prior to today's audience polarization, "when programing choices were based on garnering the largest possible number of viewers from the mass audience, the goal was to offend the fewest [and] to program the least objectionable content" (Berry & Sobieraj, 2014, p. 17). Thus, conversational restraint is often not the result of enforced rules of conduct, but of the social context and economic incentives in communication.

For deliberative theorists, liberal conversational restraint is too restrictive, as deep disagreement cannot be processed by focusing on what we agree on, but only by opening up to and assessing the weight of opposing arguments (Gutman & Thompson, 2009). Yet, such openness necessitates principles of accommodation (Gutman & Thompson, 2009), meaning that mutual respect needs to be fostered for opposing positions—requiring a moral economy. As attacks make the insulted close off against arguments and desensitize ourselves to the positions of the insulted, condemning opponents should be avoided. However, strict civility rules can maintain inequalities, as powerful groups are able to ignore civil pleas for justice and condemn uncivil ones (Huspek, 2007). Still, instead of discarding civility norms, Estlund (2008) advocates deliberative standards as a "breakdown theory": If deliberative equilibrium is broken (e.g., by power imbalances), deviations from such standards (e.g., through incivility) should be allowed, *provided* they serve to restore the balance. In this understanding, civility norms serve to prevent closed-mindedness and should be relaxed only to the extent that they do not serve this purpose.

Finally, the tradition of agonistic pluralism is often portrayed as laissez-faire toward incivility, but this only applies conditionally: While passionate, impolite and disruptive speech should be appreciated from this perspective (Mouffe, 2013), agonistic pluralism aims to foster agonistic respect, that is, a "reciprocal commitment to inject generosity and forbearance into public negotiations between parties who acknowledge that the deepest wellsprings of human inspiration are to date susceptible to multiple interpretations" (Connolly, 2005, p. 125). Thus, while agonistic theorists warn us not to be "fainthearted" in relation to uncivil talk (Laclau, 2007, p. 250), analyzing and criticizing the exclusionary effects of certain types of incivility is key.

What unites normative theories' concern for civility, then, is the telos of preventing what Medina (2013) refers to as blindness or insensitivity for the perspective of others: Certain speech acts impede constructive democratic debate because they carry a disregard for the positions of fellow debaters based on the presumption that their experiences are irrelevant or untrustworthy. Not all normative traditions would expect public debaters to understand each other, nor public debates to be frictionless, polite or

friendly. But we should expect an appreciation for both common ground and differences and, conversely, be interested in how certain forms of incivility and impoliteness, which we refer to as toxic outrage, prevent such open-mindedness.

What is Toxic Outrage?

Several studies have disentangled incivility further: Papacharissi (2004) distinguishes between impoliteness and incivility. The former denotes speech acts that are characterized by inappropriate manners such as name-calling or vulgarity, whereas the latter violates democratic principles, for example by stereotyping or calling to remove other people's rights. Muddiman (2017) refers to this as personal- versus public-level incivility. Rossini (2020), in turn, separates incivility from intolerance, with the former being a violation of common politeness norms and the latter expressing a fundamentally discriminatory intent toward people or groups based on their personal characteristics, preferences or social status. While impoliteness can also be a constructive component of online discussions, both Papacharissi (2004) and Rossini (2020) argue that democratic (public-level) incivility or intolerance is detrimental to democratic debate.

Expanding on these conceptual distinctions, what separates unacceptable from constructive forms of incivility is that they carry and create a fundamental insensitivity toward other perspectives. Toxic outrage is a rhetorical strategy that aims to foster such disregard by provoking negative emotional reactions in the audience against political opponents (Sobieraj & Berry, 2011). It contains elements of both impoliteness and democratic incivility/intolerance (Papacharissi, 2004; Rossini, 2020), but only those that are detrimental to public debate. What distinguishes toxic outrage from other forms of incivility and impoliteness are "the elements of malfeasant inaccuracy and [the strategic] intent to diminish" (Sobieraj & Berry, 2011, p. 20). While impolite, even some forms of uncivil rhetoric sometimes benefit the discussion, toxic outrage further polarizes debates (Anderson et al., 2014) by increasing closed-mindedness in the audience through moral indignation (Hwang et al., 2018). Thus, toxic "outrage is incivility writ large. It is by definition uncivil, but not all incivility is outrage" (Sobieraj & Berry, 2011, p. 20).

In an increasingly diverse digital media environment, some actors decidedly support outrageous debate by creating discussion spaces that attract this rhetoric. A widespread toxic "outrage industry" (Berry & Sobieraj, 2014) has emerged that deliberately enrages different parts of the public against each other. In this context, it has been argued that the spread of toxic outrage is linked to national political and platform-related factors (Sobieraj & Berry, 2011). By systematically analyzing toxic outrage online in majoritarian versus consensus-oriented democracies, in arenas that separate versus collapse public and private contexts and in forums that are used primarily for issue- rather than preference-driven debates, the present study sets out to test this assumption. In comparing two cases in which the respective explanatory factor is present with two cases in which the factor is absent, respectively, the research relies on the inferential "method of difference" (Mill, 1843). Thereby, looking at two cases per

group facilitates the interpretation of structural influences rather than the exploration of individual country- or arena-related idiosyncrasies. The rationale for our hypotheses is set out below.

Toxic Outrage Across Democratic Political Systems

This study relies on Lijphart's (2012) distinction of majoritarian versus consensus-oriented democracies to explain national differences in the level of toxic outrage online. While more disaggregated multi-index conceptions of democracy like *Varieties of Democracy* would facilitate the consideration of gradual variations in democratic performance, Lijphart's general typology is particularly useful for our research because it "provide[s] a rough empirical estimate of a complex and multivalent concept" (Coppedge et al., 2011, p. 252). Abstracting the tangled political architectures of the studied countries allows us to investigate cross-national structural influences *alongside* platform-related antecedents of toxic outrage online in our unique endeavor to examine these explanatory factors together. Furthermore, as the following rationale will show, the distinction of majoritarian versus consensus-oriented democracies is closely related to the polarization versus moderation of public debates and thus tightly linked to our research question. The limitations of Lijphart's typology will be reflected upon in the discussion section.

While majoritarian democracies are dominated by two competing party blocs and concentrate executive power with the majority party, consensus-oriented democracies share governing authority among multiple parties and thus focus on political compromise (Lijphart, 2012). Accordingly, in public debates, political actors strive to accommodate different perspectives in consensus-oriented democratic systems, whereas they clearly dissociate from each other in majoritarian democracies (Steiner et al., 2004). This habit to dissociate in two-party systems has further intensified in recent years, as political parties increasingly polarize by moving toward ideological extremes (Levendusky, 2009) and party elites fuel a strong "us versus them" dichotomy in public communication (McCoy & Somer, 2019). As majority parties increasingly divide into opposed camps, "they also increasingly perceive politics as a zero-sum competition, in which a win for one side is inherently a loss for the other" (Mason, 2018, p. 60).

The political system patterns of accommodation and dissociation in the different types of democracy are transferred to the electorate predominantly via the media (Levendusky, 2009). Specifically, as the media report on political issues in majoritarian democracies, they "communicate to the public the degree to which politicians are polarized along party lines" (Arceneaux & Johnson, 2015, p. 309), which, in turn, causes citizens to align more clearly with party ideologies, a process referred to as "partisan sorting" (Levendusky, 2009). While consensus-oriented democracies tend to have comparatively regulated democratic corporatist media systems, majoritarian democracies usually have rather liberal media systems (Hallin & Mancini, 2004) whose market-oriented structures further facilitate the depiction of "politics as a struggle between irreconcilably opposed parties" (Tucker et al., 2018, p. 40). As they follow the polarization trend within the political system by catering to increasingly

divided audiences, these systems can become rather polarized liberal media systems (Nechushtai, 2018) that facilitate a rise in toxic outrage spurred by (partisan) media (Berry & Sobieraj, 2014). Still, in addition to their mediated dissemination, elite polarization cues can also spread through other channels such as interpersonal encounters (Druckman et al., 2018) or social movements (Levendusky, 2009).

Indeed, studying twenty democracies, Gidron et al. (2020) show that citizens who identify with a specific party "in countries with majoritarian, single-winner voting systems tend to dislike opposition parties more intensely . . . than do partisans in countries with proportional voting systems" (p. 10). Notably, this affective polarization is not necessarily accompanied by more extreme policy positions. Mason (2015) finds that behavioral and issue position polarization are rather distinct—and that while partisan sorting contributes strongly to the former, it does not increase issue position extremity to the same extent. Political polarization in the electorate is thus primarily based on partisan identities and group membership, taking the form of an *uncivil agreement*, as part of which citizens "agree on most issues but are nevertheless growing increasingly biased, active and angry" (Mason, 2013, p. 141). Based on the aforementioned theoretical and empirical insights, we assume that citizens' anger and affective polarization manifest in user-generated discussions and therefore hypothesize:

> *H1*: The level of toxic outrage in online user comments is higher in majoritarian than in consensus-oriented democracies.

Empirically, Lijphart (2012) distinguishes majoritarian and consensus-oriented democracies along an executives-parties and a federal-unitary dimension, consisting of five indicators each. The former considers the effective number of parliamentary parties, the concentration of power in cabinet, the dominance of the executive vis-à-vis the legislative, the disproportionality and type of the electoral system and interest group pluralism in a democracy. The latter maps indicators of "[power] diffusion by means of institutional separation" (Lijphart, 2012, p. 4) like the dispersion of power on different government levels or bicameralism. Our country selection focused strongly on the executives-parties dimension while keeping the federal-unitary dimension rather constant. The previous theoretical considerations showed that the distinction of one-party majority versus multiparty coalition governments, that is, the concentration versus sharing of executive power, and the related tendency for dissociation versus accommodation in public debates, is particularly consequential for toxic outrage online. Coincidingly, the distinction of one-party majority versus multiparty coalition governments constitutes the "most important and typical difference between the two models of democracy" (Lijphart, 2012, p. 60) as well as the empirically most defining factor of the executives-parties dimension. While aiming to include nations from several continents, we also sought to keep the number of languages low to mitigate their influence in the automated analysis. This resulted in selecting the majoritarian democracies of Australia and the United States and the consensus-oriented democracies of Germany and Switzerland (where we focus on German-language debates). While all four countries have attenuated in their degree of majority- versus consensus-orientation over the last decades, they continuously

belong to the respective general types. The Supplemental Appendix A depicts the countries on Lijphart's (2012) two-dimensional map of democracy and shows their performance on the sub-dimensions of the executives-parties dimension.

Toxic Outrage Across Online Discussion Arenas

On the platform level, specific socio-technical affordances (Marwick, 2018) shape the debates in different online discussion forums. In accordance with a platform's technical design, users have distinct "perceptions of what actions are available to them [in these arenas]" (Nagy & Neff, 2015, p. 5), which frame how audiences predominantly use a certain communication space. We suggest that the perceived degree of context collapse in and the primary use function of a discussion arena may be particularly consequential for the level of toxic outrage in user debates.

Context collapse. Democratic discourse is often conceptualized as a semi-autonomous civic sphere that is decoupled from private, essentially sociable conversations to facilitate substantive contestation (Schudson, 1997). In online forums, however, a user's varying audiences often integrate into one indistinguishable collective (Vitak, 2012), with public and private contexts increasingly blurring. As discussion spaces connect family, friends, co-workers, and other acquaintances (Boyd, 2011), it can be difficult for users to determine which style of communication is socially appropriate (Rowe, 2015b). While this is of course also shaped by individual conduct, overall, the literature suggests that certain platforms afford a much stronger degree of context collapse to their users than others. On Facebook, public and private spheres mix rather strongly, as users perceive a "high salience of invisible audiences and collapsed contexts" (Rowe, 2015b, p. 543) in this environment. Even when users post in seemingly public arenas, such as on the pages of media outlets or political groups, where comments are directed primarily to unknown co-debaters, these posts are at least potentially visible to the poster's entire friend network (Hughes et al., 2012). Twitter and the website comment sections of mainstream news media, in contrast, are more public in nature. Both connect individuals to strangers more often than Facebook, "focus[ing] less on 'who you are' and more on what you have to say" (Hughes et al., 2012, p. 562; Rossini, 2020). The perceived degree of context collapse is thus weaker in these forums.

Online arenas that separate public from private contexts tend to be characterized by lower levels of identifiability. This reduces the risk of being held accountable for one's statements and encourages various forms of incivility (Santana, 2014). Rowe (2015a), for example, found that user posts contain more impoliteness and democratic incivility in the Washington Post's website comment section than on the paper's public Facebook page. Similarly, user comments on the more de-individuated Youtube channel of the White House have been found to be more impolite than those on the government's Facebook page (Halpern & Gibbs, 2013). We thus hypothesize:

> *H2*: The level of toxic outrage in online user comments is higher in arenas that separate public and private contexts more clearly than in those that mix the two.

Table 1. Communication Arenas by Degree of Context Collapse and Primary Use Function.

	Primary use function	
Context collapse	Issue-driven, pluralistic discussion	Preference-driven, like-minded discussion
Weak (rather public context)	Mainstream news media website comment sections	Twitter
Strong (mixed public and private context)	Mainstream news media Facebook pages	Facebook pages of partisan collective actors and alternative media with a stance

Primary use function. Another core idea of democratic theory is that public contestation should take place across lines of difference (Gutman & Thompson, 2009). However, instead of engaging with diverse views, online political discussions are often rather polarized (Yarchi et al., 2021). The primary use function of a discussion arena indicates whether this forum is used by individuals primarily for issue-driven debates that evolve pluralistically around a contested issue or rather to conduct preference-driven discussions that bring together like-minded people.

In the political context, Twitter mostly affords rather preference-driven debates with ingroup-oriented structures (Freelon, 2015; Yarchi et al., 2021), in which individuals engage with contents (Himelboim et al., 2013) and users (Vaccari et al., 2016) of similar political preferences. While hashtags could potentially bring together individuals with different views, in reality, they primarily integrate those with similar positions. Likewise, research suggests that the Facebook pages of partisan collective actors and alternative media are used primarily for discussions among like-minded people (Maia & Rezende, 2016; Maia et al., 2021). In contrast, the website comment sections and Facebook pages of mainstream news media assemble a readership base that is connected by an interest in the topic of the original article (Freelon, 2015) and whose political views have been shown to be rather diverse (Nelson & Webster, 2017). By investigating two different kinds of discussion arenas on Facebook, we account for the fact that the platform's socio-technical affordances encourage different primary use functions.

Research shows that heterogeneous discussion spaces are more prone than homogeneous forums to foster disrespectful behavior (Maia & Rezende, 2016). Insults, for example, were found to be much more common in user comments on news websites than in debates on Twitter, which include more ingroup-oriented elements (Freelon, 2015). We thus hypothesize:

> *H3*: The level of toxic outrage in online user comments is higher in arenas that are used primarily for issue-driven debates with plural opinions than in forums that afford rather preference-driven, like-minded discussions.

Table 1 classifies the communication arenas analyzed in this study according to their degree of context collapse and primary use function.

Table 2. Overview of Data Analyzed.

	Country				
Communication arena	Australia	United States	Germany	Switzerland	Total
News website comment sections	5,054	15,850	6,133	3,306	30,343
Mainstream media Facebook pages	4,527	44,190	4,753	760	54,230
Facebook pages of partisan actors and alternative media	30,733	130,400	12,458	3,069	176,660
Twitter	22,528	771,054	176,478	5,258	975,318
Total	62,842	961,494	199,822	12,393	1,236,551

Methodology

In an automated content analysis, this study investigated 1,236,551 user contributions on the public role of religion and secularism in society, published from August 2015 to July 2016. Table 2 shows the data analyzed per discussion arena per country.

Case Study

Debates on the public role of religion and secularism in society are a hard case for civil contestation because religiously grounded value systems can exhibit elements of fundamentalism that make public discussions comparatively closed and apodictic. In religiously tainted debates, some speakers might more readily construe their opponents as enemies and denigrate their views. In the run up to the 2016 Australian and US elections and at the height of the European refugee movement 2015/2016, the public role of religion and secularism in society was hotly debated in all four countries. Amid rising skepticism about the immigration of religious minorities into Western democracies, the material we study mirrors quarrels on the ensuing expectations for cultural adaption, such as the wearing of headscarves in public, as well as longstanding issues in which religious and secular camps are divided, such as abortion or same-sex marriage.

Data Collection

Data collection took place in a carefully validated multi-step process warranting data comparability. We systematically collected material of users commenting on similar issues and positions in all four communication arenas and all four countries. This collection of user comments used a diligently selected pool of news articles and blog posts as its starting point and branched out into four data collection paths, one for each discussion arena. Accordingly, data collection started with a dataset of 1,127 news

articles and blog posts on the subject of interest, issued from August 2015 to July 2016 by leading print newspapers, news websites and political blogs (Supplemental Appendix B) in the four countries. The studied outlets are the leading outlets of record in each category in the respective societies, according to 16 or more academic experts we surveyed in each country. Likewise, to build this corpus, an expert survey was conducted among 76 communication and religious studies scholars in the societies, who named relevant debates on the public role of religion and secularism in the respective society and a list of keywords associated with each debate. Based on these keywords, the articles and blog posts were selected in a novel, expert-informed topic modeling process (for detailed information see Rinke et al., 2021). The base corpus and the expert survey results then guided the following four data collection paths:

1. For contributions from mainstream news media websites (Supplemental Appendix B/B-1), we identified all news website articles from the base corpus featuring a user comment section (115 out of the total of 400 news website articles in the base corpus) and collected the posts therein.
2. For comments from mainstream news media's Facebook pages (Supplemental Appendix B/B-2), we identified all news website articles from the base corpus that had been posted on the respective media outlet's Facebook page (76 out of the total of 400 news website articles in the base corpus) and collected all comments below these.
3. To collect contributions from the Facebook pages of partisan collective actors and alternative media, we first identified relevant pages by drawing on a list of all actors mentioned in the base corpus. Those collective actors and alternative media with a particular interest in the public role of religion and secularism in society and an active Facebook page in the period of investigation (e.g., the *Secular Coalition for America* and *Christianity Today*) were chosen for analysis. As these actors were referred to by the leading print newspapers, news websites and political blogs in the studied countries, they can also be regarded to be among the most relevant of their kind in each of these societies—which makes a country comparison possible. Facebook's "similar page" function, desktop research and consulting selected academic experts all served to expand and substantiate the selection. In total, 76 Facebook pages of partisan collective actors and 41 Facebook pages of alternative media were selected for analysis (Supplemental Appendix B/B-3). We collected all entries posted by these pages in the period of investigation and scored them for subject relevance with topic models that were built from extensive text corpora (Rinke et al., 2021) and that relied on the expert survey keywords. A cut-off for relevance was defined with gold standards of $n = 300$ comments in each country, each of which was scored by two trained coders with Krippendorff's $\alpha_{nominal}$ of .78. This resulted in 4,899 relevant Facebook seed posts from partisan collective actors and alternative media for which all user comments were collected.
4. To identify tweets, we researched all available Twitter profiles of the partisan collective actors and alternative media identified in the previous step and

created a list of their 5,000 most frequently mentioned hashtags from August 2015 to July 2016. Inspired by this list, we selected 64 Twitter debate hashtags (Supplemental Appendix B/B-4) that strongly related to at least one of the debates on the public role of religion and secularism in society named in the expert survey. In the United States, for instance, the debate on religious opposition against same-sex marriage being prominently referred to in the survey led to the selection of #kimdavis. All tweets that featured at least one of the hashtags in the period of investigation were collected.

The data was collected for a large-scale research program that examines the democratic quality of user-generated debates comparatively. A subsample of the data analyzed in this study was therefore also used in a previously published study on the integrative complexity of online user comments across different types of democracy and discussion arenas (Jakob et al., 2021). However, each of these investigations focuses on a distinct dimension of debate quality, and thus makes a unique contribution to different lines of research. While this study centers on toxic outrage as a violation of civility norms in online debates, that is, on a rather sentiment-based construct, the study on integrative complexity concentrates on the argumentative quality of user comments online, that is, on a more substantive, content-related dimension of debate quality.

Automated Content Analysis

We combined an off-the-shelf dictionary with machine learning to measure toxic outrage in the collected posts—a novel automated method suggested by Dobbrick et al. (2021). This leveraged the knowledge incorporated in the word list and tailored it to our needs, thus reducing the amount of hand-coded data required for stand-alone machine learning. Aligning with CRISP-DM (Shearer, 2000), our approach followed five steps: Generating a gold standard, pre-processing it by applying LIWC, then training, evaluating and deploying the machine learning model. Since our instrument cannot classify visual material, we focused on the text of the posts.

Generating the gold standard. Following Sobieraj and Berry (2011), toxic outrage as an effort to cause a negative emotional reaction in the audience may be elicited by 13 rhetorical means, including "insulting language, name calling, emotional display, emotional language, verbal fighting/sparring, character assassination, misrepresentative exaggeration, mockery, conflagration, ideologically extremizing language, slippery slope, belittling, and obscene language" (p. 26). Based on the authors' category descriptions (Sobieraj & Berry, 2011), two individuals were trained to code the gold standard. The unit of analysis was the comment. Importantly, rather than coding individual forms of incivility, to facilitate the automated measurement, toxic outrage was coded as a binary variable in this study, that is, as either present or absent in a comment. Thereby, toxic outrage was present if at least one of the 13 modes of outrage occurred in a post. In a pretest on 320 randomly selected user comments (20 per arena

per country), Krippendorff's $\alpha_{nominal}$ was .80. In the main coding, 200 posts from each arena in each country were scored, that is, 3,200 comments. Each item was assessed by both coders, with disagreements resolved consensually.

By using this pre-defined concept of toxic outrage, this study focused on investigating "researcher-defined uncivil content" (Van Duyn & Muddiman, 2020, p. 12). Based on the gold standard, the automated classifier identified a set of English and German words, respectively, that best predict toxic outrage across all discussion arenas in the majoritarian versus consensus-oriented democracies. The advantage of this ex-ante conceptualization is that it enables a highly systematic comparison of the prevalence of toxic outrage in different types of democracy and arenas. The analysis cannot, however, provide insights into how users *perceive* this incivility, which can likewise vary across individuals, social contexts, and countries (Kenski et al., 2020).

Pre-processing with off-the-shelf dictionary. The gold standard was pre-processed by applying all LIWC2015/DE-LIWC2015 categories (Pennebaker et al., 2015) to the English and German posts, respectively. This includes linguistic and formal features like word count or the share of words longer than six letters.

Training the machine learning model. An M5P machine learning algorithm (Quinlan, 1992; Wang & Witten, 1996) was then trained on the gold standard. The model combines decision tree with regression analysis. It splits the data into subsets so that the variance of the target variable across the instances is minimized. When the number of data points or their variance in the subsets is below a certain threshold, it stops. Finally, a linear regression model is fit to predict the outcome in that subset. The tree-based M5P automatically deals with variable selection, variable importance, missing values, normalization, and variable interactions (Song & Lu, 2015). It thus works well with LIWC-generated features, as many dictionary categories are combinations of others, and hence highly correlated, and they vary between the English and the German LIWC. Combining decision trees with linear regression, M5P predicts continuous values that need to be discretized for binary measurement—in our case with the split point generated automatically using the gold standard.

Evaluation. We assessed the performance of our approach with a 10-fold cross-validation on the gold standard. This trains the model on nine equal folds of data and holds one out for evaluation. The process is repeated ten times, so every subsample is used as the validation set once. The performance is then averaged over the ten runs. Table 3 shows the performance metrics, indicating that our instrument works comparatively well (see Supplemental Appendix C/Table 1 for performance per arena and country).

Applying the model to the corpus. Finally, we pre-processed the full data corpus in the same way as the gold standard, applied the trained M5P model to predict toxic outrage for each comment in the set and discretized the score for statistical analysis. The workflow is available in Supplemental Appendix C.

Table 3. Average Performance of LIWC and M5P Based Outrage Classification on Gold Standard.

	Outrage classification performance							
Approach	TN	FP	FN	TP	Accuracy	Precision	Recall	F1
LIWC & M5P	2,349	135	268	648	.80	.88	.71	.76

Note. Performance calculated on manually annotated gold standard of $N = 3,200$ comments (200 contributions from each arena in each country). TN = true negatives; FP = false positives; FN = false negatives; TP = true positives.

Results

Overall, 17.67% of the user contributions in our dataset of $N = 1,236,551$ contained toxic outrage. Figure 1 features the share of toxic outrage per country and communication arena. It shows that toxic outrage was more frequent in some countries and forums than others, with patterns that lend initial support for the formulated hypotheses.

To test these hypotheses, we ran binominal logistic regression analysis. It estimates the logistic transformation of the probability of an event, that is, the log odds of toxic outrage to be present in a comment. Specifically, the presented models estimate the log odds of toxic outrage to occur in a post in a specific group versus the log odds of toxic outrage to occur in a post in the reference group (i.e., in majoritarian vs. consensus-oriented democracies, under separated vs. collapsed contexts, in issue-driven vs. preference-driven arenas). To avoid dominant effects of individual countries or arenas that stem from the varying amounts of collected comments, the data was balanced by overweighting the arenas with lower case numbers in all countries according to the size of the arena in which the largest number of comments was collected (Supplemental Appendix D/D-1). A robustness check showed that the hypothesized main effects for the democratic system, context collapse and primary use function (H1–H3) are substantively the same when tested on the weighted sample versus the hand-coded, stratified randomly sampled gold standard—and that the effect sizes fall within the confidence intervals of a bootstrapped regression analysis on a stratified undersampled dataset (Supplemental Appendix D/D-2). This demonstrates that the weighting did not affect our general findings. Since longer posts are more likely to contain toxic outrage, the regression analysis controlled for comment length. To ease interpretation, this covariate was log-transformed to reduce skewness[1] and mean-centered, so that the intercept is the expected value of Y when comment length is set to the mean instead of zero. Tables 4 and 5 report the coefficients of two separate regression models that were computed to test the hypotheses and to conduct a more detailed country and arena comparison.

Table 4 shows that the data backed our first hypothesis, which stated that the level of toxic outrage in online user comments is higher in majoritarian than in consensus-oriented democracies. The predicted chance of a comment to contain toxic outrage

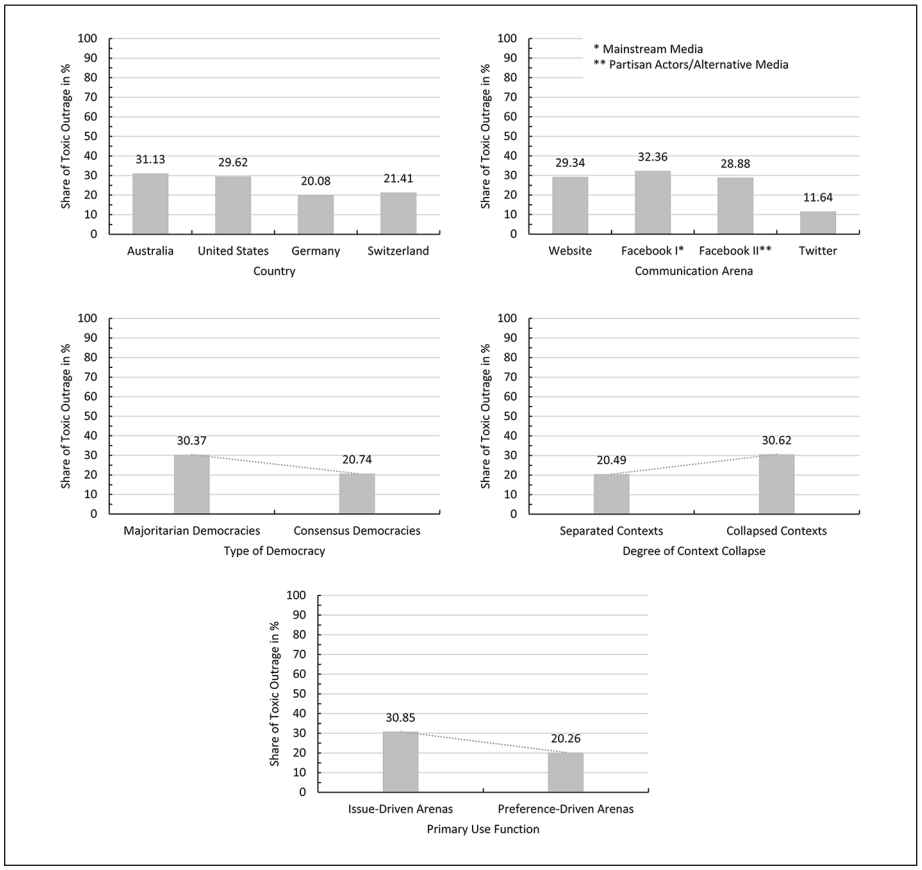

Figure 1. Share of user comments that contain toxic outrage in percent ($N = 1{,}236{,}551$).

rose from 15.9% in consensus-oriented political systems ($SE < 0.001$, $p < .001$, 95% CI [.159, .159]) to 30.5% in majoritarian democracies ($SE < 0.001$, $p < .001$, 95% CI [.305, .306]). Yet, this possibility differed slightly between both Switzerland (16.2%, $SE < 0.001$, $p < .001$, 95% CI [.161, .162]) and Germany (15.5%, $SE < 0.001$, $p < .001$, 95% CI [.155, .156]) and Australia (32.5%, $SE < 0.001$, $p < .001$, 95% CI [.324, .326]) and the United States (28.5%, $SE < 0.001$, $p < .001$, 95% CI [.285, .286]) (see also Table 5).

However, contrary to our assumption, the level of toxic outrage in online user comments was lower in arenas that separate public and private contexts more clearly (Twitter and news website comment sections) than in those mixing the two (Facebook), as shown in Table 4. The regression predicted that while a user comment has a 14.6% chance to carry toxic outrage under separated contexts ($SE < 0.001$, $p < .001$, 95% CI [.146, .146]), it is 32.7% under collapsed contexts ($SE < 0.001$, $p < .001$, 95% CI

Table 4. Binomial Logistic Regression Model for Hypotheses Testing.

					95% CI	
Independent variable	*b*	*p*	*SE*	*OR*	*LL*	*UL*
H1: Political system						
Majoritarian (vs. consensus)	.572	<.001	0.003	1.772	1.762	1.781
H2: Context collapse						
Separated contexts (vs. collapsed)	−1.756	<.001	0.004	0.173	0.171	0.174
H3: Primary use function						
Issue-driven (vs. preference-driven)	.214	<.001	0.003	1.239	1.233	1.245
Interaction effects						
Separated contexts × issue-driven	.742	<.001	0.005	2.099	2.080	2.119
Separated contexts × majoritarian	.614	<.001	0.005	1.847	1.830	1.865
Issue-driven × majoritarian	−.140	<.001	0.004	0.869	0.863	0.876
Separated contexts × issue-driven × majoritarian	.141	<.001	0.006	1.151	1.138	1.165
Comment length (characters, log transformed/mean-centered)	.480	<.001	0.001	1.617	1.615	1.619
Constant	−1.080	<.001	0.002	0.340	0.338	0.341
$\chi^2(8)$ = 1,376,597, $p < .0001$						
−2*log likelihood = 12,647,365, R^2 = .106 (Cox & Snell), .155 (Nagelkerke)						
Weighted *N* = 12,336,864						

Note. Tables 4 and 5 report separate regression models. OR = odds ratio; LL = lower limit; UL = upper limit.

[.327, .327]). Our second hypothesis was thus rejected based on an effect into the opposite direction.

Our third hypothesis, again, was supported. As Table 4 shows, the level of toxic outrage in online user comments was higher in arenas that are used primarily for issue-driven debates with plural opinions than in forums that afford rather preference-driven, like-minded discussions. The probability that a user post contains toxic outrage was predicted at 27.5% ($SE < 0.001, p < .001$, 95% CI [.275, .276]) and 18.0% ($SE < 0.001$, $p < .001$, 95% CI [.179, .180]), respectively. In a more detailed breakdown, Table 5 shows that in relation to Twitter, toxic outrage was increasingly more likely in news website comment sections, on the Facebook pages of partisan collective actors and alternative media and on the Facebook pages of mainstream news media, with the chance of a user comment to contain toxic outrage at 9.6% ($SE < 0.001, p < .001$, 95% CI [.095, .096]), 21.6% ($SE < 0.001, p < .001$, 95% CI [.215, .216]), 31.0% ($SE < 0.001$, $p < .001$, 95% CI [.310, .311]), and 34.3% ($SE < 0.001, p < .001$, 95% CI [.343, .344]), respectively.

Apart from the hypothesized main effects, as Tables 4 and 5 show, there were significant interaction effects between the three predictor variables. These mainly stemmed from the fact that relative to the discovered country and platform effects, the chance of

Table 5. Binomial Logistic Regression Model for Country and Arena Analysis.

					95% CI	
Independent variable	*b*	*p*	*SE*	*OR*	*LL*	*UL*
Country (vs. Switzerland)						
Germany	.253	<.001	0.007	1.288	1.271	1.305
United States	1.216	<.001	0.006	3.373	3.334	3.412
Australia	1.412	<.001	0.006	4.106	4.060	4.153
Arena (vs. Twitter)						
News websites	1.258	<.001	0.006	3.517	3.478	3.557
Facebook pages partisan collective actors/alternative media	1.873	<.001	0.006	6.505	6.432	6.578
Facebook pages mainstream news media	2.160	<.001	0.006	8.673	8.579	8.769
Interaction effects						
Germany × News websites	−.608	<.001	0.008	0.544	0.536	0.553
Germany × Facebook pages partisan actors / alternative media	−.223	<.001	0.008	0.800	0.788	0.813
Germany × Facebook pages mainstream news media	−.371	<.001	0.008	0.690	0.680	0.700
United States × News websites	−.247	<.001	0.007	0.781	0.771	0.792
United States × Facebook pages partisan actors / alternative media	−.838	<.001	0.007	0.433	0.427	0.439
United States × Facebook pages mainstream news media	−.867	<.001	0.007	0.420	0.415	0.426
Australia × News websites	−.349	<.001	0.007	0.705	0.696	0.715
Australia × Facebook pages partisan actors / alternative media	−.638	<.001	0.007	0.528	0.521	0.536
Australia × Facebook pages mainstream news media	−1.005	<.001	0.007	0.366	0.361	0.371
Comment length (characters, log transformed/mean-centered)	.482	<.001	0.001	1.619	1.616	1.621
Constant	−2.968	<.001	0.005	0.051	0.051	0.052
$\chi^2(16) = 1{,}401{,}932$, $p < .0001$						
−2*log likelihood = 12,622,030, $R^2 = .107$ (Cox & Snell), .158 (Nagelkerke)						
Weighted $N = 12{,}336{,}864$						

Note. Tables 4 and 5 report separate regression models. OR = odds ratio; LL = lower limit; UL = upper limit.

a user comment to contain toxic outrage in the news website comment sections of mainstream media was comparatively higher in the United States than in the other three countries and comparatively lower in Germany. Similarly, the predicted probability of

a user post to carry toxic outrage on the Facebook pages of partisan actors and alternative media was comparatively higher in Australia than in the other countries (see Supplemental Appendix D/D-4 for a graphical display of the interactions).

Discussion

This study was the first to investigate national political and platform-related context factors of toxic outrage, that is, destructive incivility, in online discussions together. It analyzed user comments from Australia, the United States, Germany and Switzerland, comparing posts from the website comment sections and Facebook pages of mainstream news media, the Facebook pages of partisan collective actors and alternative media as well as from Twitter.

The level of toxic outrage in online user comments was higher in majoritarian than in consensus-oriented democracies. This suggests that the civil "spirit of accommodation" (Lijphart, 1975, p. 103) typical for consensus-oriented democracies and the tendency for dissociation and polarization in majoritarian democracies translate from elite discourse into user-generated debates, where they facilitate more or less constructive engagement. Political system characteristics may thus not only serve as incentive structures for political actors in public debates but also as modeling patterns for citizens. This supports deliberative democrats' case for the advantages of consensus-oriented democracies, which, they argue, can incentivize social learning across political camps and thus help mitigate deep societal conflicts (Dryzek, 2005).

However, the level of toxic outrage in online user comments differed slightly between countries of the same democratic system. In explaining these nuances, this study is constrained by the limits of Lijphart's (2012) general distinction between majoritarian and consensus-oriented democracies. While this typology allowed us to study cross-national influences *alongside* platform-related antecedents of toxic outrage online, it prevents us from pinpointing exactly which (combination of) characteristics of majoritarian and consensus-oriented democratic systems elicit the difference in destructive incivility (Coppedge et al., 2011). The type and strength of majoritarianism or consensus-orientation could be of interest in this respect. For instance, when minority parties gain the parliamentary majority in disproportional majoritarian political systems (as the Republican party did in the United States), this could drive political polarization even further (McCoy & Somer, 2019). Similarly, while we theorize the process of how the political system influences citizen-to-citizen interactions online, our study does not directly observe the intervening stages of this process. In addition, Lijphart's typology may suffer from multicollinearity, with confounding factors like "cultural norms, historical pathways and [other] contextual circumstances" (Bormann, 2010, p. 6) also at play. The regulations that governments issue to mitigate uncivil conduct online or the degree to which the legislature sanctions this behavior could, for example, be especially relevant in this regard. Future research should focus more explicitly on explaining country-level differences in toxic outrage online by relying on a larger number of countries and more fine-grained empirical notions of democracy to disentangle the various factors at play.

Contrary to our expectation, the level of toxic outrage in online user comments was lower in arenas that separate public and private contexts more clearly than in those that collapse varying audiences into one group. Thus, after all, online debates were more likely to be constructive when they were conducted in semi-autonomous spheres of democratic engagement that are decoupled from private life more distinctly (Schudson, 1997). While this contrasts prior research showing that lower identifiability encourages incivility online (Halpern & Gibbs, 2013; Rowe, 2015a; Santana, 2014), it mirrors a study by Rossini (2019). Comparing comments on news media websites and Facebook pages in Brazil, she, too, finds that the identifiability of users and the ensuing social constraints on Facebook do not prevent citizens from being uncivil. The platform's "users may be interacting with others outside of their networks when commenting on news stories and therefore might not feel as constrained by their social ties to adopt uncivil rhetoric" (Rossini, 2019, p. 236). The attitude with which citizens enter discussions in different arenas may be an additional explanation for this. When debating under separated contexts, users seek these exchanges rather actively (Springer et al., 2015). Under collapsed contexts on Facebook, in contrast, they may react more spontaneously to public content that appears on their timeline and appeals to them emotionally. Individuals discussing under separated contexts might thus be better prepared to control their behavior and refrain from toxic outrage. Theoretically, this finding overlaps with the liberal democratic understanding that the separation of public and private spheres mitigates democratic conflict. If individuals bracket their private convictions and relationships from public debates, liberal theory argues, they are more readily able to respect opposing views within the public sphere (Ackerman, 1989; Holmes, 1988).

In relation to the primary use function, we found that the level of toxic outrage in online user comments was higher in arenas that are used primarily for issue-driven debates with plural opinions than in forums that afford rather preference-driven, like-minded discussions. As hypothesized, this supports research "indicat[ing] that people are more motivated to use foul language when interacting with those who hold different views" (Maia & Rezende, 2016, p. 129). This is an important reminder that the mere confrontation with opposing positions in online discussions does not automatically lead to more constructive engagement but may in fact fuel hostility and opinion polarization. Likely, hearing different perspectives only fosters more civil debate when it is coupled with "apophatic listening" (Dobson, 2014) that is aimed at a clear understanding of what the other wants to say. To some degree, this counters agonistic theorist's implicit assumption that public contestation as such, through the mere exposure to opposing views, can foster agonistic respect (Mouffe, 2013). At the same time, agonistic theorists would remind us to supplement this finding in future research by studying constructive siblings of toxic outrage, that is, passionate, sometimes impolite rhetoric which could make a beneficial contribution to online discussions (Jamieson et al., 2017).

Relative to the discovered country and platform effects, toxic outrage was comparatively more likely in news website comment sections in the United States and comparatively less likely in news website comment sections in Germany. This could

be due to different content moderation styles in these countries, that may either reduce or encourage online incivility further (Ziegele et al., 2018). Moreover, the chance of a user comment to contain toxic outrage on the Facebook pages of partisan actors and alternative media was comparatively higher in Australia than in the other three countries, which may indicate that these actors provide discussion spaces for particularly radicalized individuals in this country. Again, these findings set the stage for a more in-depth investigation of country-related idiosyncrasies going forward.

On the platform level, this study is limited by the fact that by focusing on two particularly consequential socio-technical affordances, it necessarily disregards others. For example, as many platform providers have recently changed their policies and infrastructures for detecting and handling hate speech, their own role in fostering a more constructive online debate culture could be a future research focus. Furthermore, substantiating the above findings by investigating additional platforms is important. This is also true more generally for the topical contexts in which online discussions take place. In addition, to the benefit of measuring toxic outrage automatically on a large scale, this study did not zoom in on the different forms of toxic outrage online (Sobieraj & Berry, 2011). A more fine-grained analysis could generate more insights into which of the rhetorical strategies involved dominate over others in the digital sphere.

Ultimately, research into toxic outrage is concerned with the factors that foster constructive and respectful democratic engagements online. Our study suggests that user-generated debate flourishes in political environments that incentivize actors to strive for compromise, put relevant issues center stage and make room for public debate at a relative distance from purely social conversation. As interactive moderation (Esau et al., 2017; Ziegele et al., 2018) may be a promising way to deal with toxic outrage online, research and practice should focus specifically on developing civic technologies that can foster more constructive online engagements across the board, beyond the simple deletion of undesired posts.

Data Availability

The data underlying this article can be shared upon reasonable request.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant number 260291564.

ORCID iD

Julia Jakob https://orcid.org/0000-0003-2340-5193

Supplemental Material

Supplemental material for this article is available online at The Open Science Framework: https://doi.org/10.17605/OSF.IO/XQ8BK.

Note

1. Since some posts contained only a visual, their written comment length was zero. Therefore, we used Laplace smoothing and added one (+1) to each comment length before log-transforming the covariate.

References

Ackerman, B. (1989). Why dialogue? *The Journal of Philosophy*, *86*(1), 5–22. https://doi.org/10.2307/2027173

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The "nasty effect": Online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication*, *19*(3), 373–387. https://doi.org/10.1111/jcc4.12009

Arceneaux, K., & Johnson, M. (2015). More a symptom than a cause: Polarization and partisan news media in America. In J. A. Thurber & A. Yoshinaka (Eds.), *American gridlock: The sources, character, and impact of political polarization* (pp. 309–336). Cambridge University Press.

Bächtiger, A., Niemeyer, S., Neblo, M., Steenbergen, M. R., & Steiner, J. (2010). Disentangling diversity in deliberative democracy: Competing theories, their blind spots and complementarities. *Journal of Political Philosophy*, *18*(1), 32–63. https://doi.org/10.1111/j.1467-9760.2009.00342.x

Berry, J. M., & Sobieraj, S. (2014). *The outrage industry: Political opinion media and the new incivility*. Oxford University Press.

Bormann, N. (2010). Patterns of democracy and its critics. *Living Reviews in Democracy*, *2*(1), 1–14. https://ethz.ch/content/dam/ethz/special-interest/gess/cis/cis-dam/CIS_DAM_2015/WorkingPapers/Living_Reviews_Democracy/Bormann.pdf

Boyd, D. (2011). Social network sites as networked publics: Affordances, dynamics, and implications. In Z. Papacharissi (Ed.), *A networked self: Identity, community, and culture on social network sites* (pp. 39–58). Routledge.

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

Connolly, W. E. (2005). *Pluralism*. Duke University Press.

Coppedge, M., Gerring, J., Altman, D., Bernhard, M., Fish, S., Hicken, A., Kroenig, M., Lindberg, S. I., McMann, K., Paxton, P., Semetko, H. A., Skaaning, S. E., Staton, J., & Teorell, J. (2011). Conceptualizing and measuring democracy: A new approach. *Perspectives on Politics*, *9*(2), 247–267. https://doi.org/10.1017/s1537592711000880

Dobbrick, T., Jakob, J., Chan, C. H., & Wessler, H. (2021). Enhancing theory-informed dictionary approaches with "glass-box" machine learning: The case of integrative complexity in social media comments. *Communication Methods and Measures.* Advance online publication. https://doi.org/10.1080/19312458.2021.1999913

Dobson, A. (2014). *Listening for democracy: Recognition, representation, reconciliation*. Oxford University Press.

Druckman, J. N., Levendusky, M. S., & McLain, A. (2018). No need to watch: How the effects of partisan media can spread via interpersonal discussions. *American Journal of Political Science*, *62*(1), 99–112. https://doi.org/10.1111/ajps.12325

Dryzek, J. S. (2005). Deliberative democracy in divided societies: Alternatives to agonism and analgesia. *Political Theory*, *33*(2), 218–242. https://www.jstor.org/stable/30038413.

Esau, K., Friess, D., & Eilders, C. (2017). Design matters! An empirical analysis of online deliberation on different news platforms. *Policy & Internet*, *9*(3), 321–342. https://doi.org/10.1002/poi3.154

Estlund, D. M. (2008). *Democratic authority: A philosophical framework*. Princeton University Press.

Freelon, D. (2015). Discourse architecture, ideology, and democratic norms in online political discussion. *New Media & Society*, *17*(5), 772–791. https://doi.org/10.1177/1461444813513259

Friess, D., & Eilders, C. (2015). A systematic review of online deliberation research. *Policy & Internet*, *7*(3), 319–339. https://doi.org/10.1002/poi3.95

Gidron, N., Adams, J., & Horne, W. (2020). *American affective polarization in comparative perspective*. Cambridge University Press.

Gutman, A., & Thompson, D. (2009). *Why deliberative democracy?* Princeton University Press.

Hallin, D. C., & Mancini, P. (2004). *Comparing media systems: Three models of media and politics*. Cambridge University Press.

Halpern, D., & Gibbs, J. (2013). Social media as a catalyst for online deliberation? Exploring the affordances of facebook and YouTube for political expression. *Computers in Human Behavior*, *29*(3), 1159–1168. https://doi.org/10.1016/j.chb.2012.10.008

Himelboim, I., McCreery, S., & Smith, M. (2013). Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication*, *18*(2), 40–60. https://doi.org/10.1111/jcc4.12001

Holmes, S. (1988). Gag rules and the politics of omission. In J. Elster & R. Slagstad (Eds.), *Constitutionalism and democracy* (pp. 19–58). Cambridge University Press.

Hughes, D. J., Rowe, M., Batey, M., & Lee, A. (2012). A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior*, *28*(2), 561–569. https://doi.org/10.1016/j.chb.2011.11.001

Huspek, M. (2007). Normative potentials of rhetorical action within deliberative democracies. *Communication Theory*, *17*(4), 356–366. https://doi.org/10.1111/j.1468-2885.2007.00302.x

Hwang, H., Kim, Y., & Kim, Y. (2018). Influence of discussion incivility on deliberation: An examination of the mediating role of moral indignation. *Communication Research*, *45*(2), 213–240. https://doi.org/10.1177/0093650215616861

Jakob, J., Dobbrick, T., & Wessler, H. (2021). The integrative complexity of online user comments across different types of democracy and discussion arenas. *The International Journal of Press/Politics*. Advance online publication. https://doi.org/10.1177/19401612211044018

Jamieson, K. H., Volinsky, A., Weitz, I., & Kenski, K. (2017). The political uses and abuses of civility and incivility. In K. Kenski & K. H. Jamieson (Eds.), *The Oxford handbook of political communication* (pp. 205–218). Oxford University Press.

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

Kies, R. (2010). *Promises and limits of web-deliberation*. Palgrave Macmillan.

Laclau, E. (2007). *On populist reason*. Verso.

Levendusky, M. (2009). *The partisan sort*. University of Chicago Press.

Lijphart, A. (1975). *The politics of accommodation: Pluralism and democracy in the Netherlands*. University of California Press.

Lijphart, A. (2012). *Patterns of democracy: Government forms and performance in thirty-six countries*. Yale University Press.

Maia, R. C. M, Hauber, G., Choucair, T., & Crepalde, N. J. (2021). What kind of disagreement favors reason-giving? Analyzing online political discussions across the broader public sphere. *Political Studies*, *69*, 108–128. https://doi.org/10.1177/0032321719894708

Maia, R. C. M., & Rezende, T. A. S. (2016). Respect and disrespect in deliberation across the networked media environment: Examining multiple paths of political talk. *Journal of Computer-Mediated Communication*, *21*(2), 121–139. https://doi.org/10.1111/jcc4.12155

Marwick, A. E. (2018). Why do people share fake news? A sociotechnical model of media effects. *Georgetown Law Technology Review*, *2*(2), 475–512. https://georgetownlawtechreview.org/why-do-people-share-fake-news-a-sociotechnical-model-of-media-effects/

Mason, L. (2013). The rise of uncivil agreement: Issue versus behavioral polarization in the American electorate. *American Behavioral Scientist*, *57*(1), 140–159. https://doi.org/10.1177/0002764212463363

Mason, L. (2015). "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science*, *59*(1), 128–145. https://doi.org/10.1111/ajps.12089

Mason, L. (2018). Losing common ground: Social sorting and polarization. *Forum*, *16*(1), 47–66. https://doi.org/10.1515/for-2018-0004

McCoy, J., & Somer, M. (2019). Toward a theory of pernicious polarization and how it harms democracies: Comparative evidence and possible remedies. *The Annals of the American Academy of Political and Social Science*, *681*(1), 234–271. https://doi.org/10.1177/0002716218818782

Medina, J. (2013). *The epistemology of resistance: Gender and racial oppression, epistemic injustice, and resistant imaginations*. Oxford University Press.

Mill, J. S. (1843). *A system of logic, ratiocinative and inductive*. Cambridge University Press.

Mouffe, C. (2013). *Agonistics: Thinking the world politically*. Verso.

Muddiman, A. (2017). Personal and public levels of political incivility. *Journal of International Communication*, *11*, 3182–3202. https://ijoc.org/index.php/ijoc/article/view/6137/2106

Nagy, P., & Neff, G. (2015). Imagined affordance: Reconstructing a keyword for communication theory. *Social Media + Society*, *1*(2), 1–9. https://doi.org/10.1177/2056305115603385

Nechushtai, E. (2018). From liberal to polarized liberal? Contemporary US news in Hallin and Mancini's typology of news systems. *The International Journal of Press/Politics*, *23*(2), 183–201. https://doi.org/10.1177/1940161218771902

Nelson, J. L., & Webster, J. G. (2017). The myth of partisan selective exposure: A portrait of the online political news audience. *Social Media + Society*, *3*(3), 1–13. https://doi.org/10.1177/2056305117729314

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

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. University of Texas at Austin. http://hdl.handle.net/2152/31333

Quinlan, J. R. (1992 , November 16–18). *Learning with continuous classes* [Conference session]. In A. Adams & L. Sterling (Eds.), Proceedings of the 5th Australian Joint Conference

on Artificial Intelligence, Hobart, Tasmania. (pp. 343–348). World Scientific Publishing. https://doi.org/10.1142/1897

Rawls, J. (1987). The idea of an overlapping consensus. *Oxford Journal of Legal Studies*, *7*(1), 1–25. https://www.jstor.org/stable/764257

Rinke, E. M., Dobbrick, T., Löb, C., Zirn, C., & Wessler, H. (2021). Expert-informed topic models for document set discovery. *Communication Methods and Measures*. Advance online publication. https://doi.org/10.1080/19312458.2021.1920008

Rossini, P. (2019). Toxic for whom? Examining the targets of uncivil and intolerant discourse in online political talk. In P. Moy & D. Matheson (Eds.), *Voices: Exploring the shifting contours of communication* (pp. 221–242). Peter Lang.

Rossini, P. (2020). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*. Advance online publication. https://doi.org/10.1177/0093650220921314

Rowe, I. (2015a). Civility 2.0: A comparative analysis of incivility in online political discussion. *Information Communication & Society*, *18*(2), 121–138. https://doi.org/10.1080/1369118x.2014.940365

Rowe, I. (2015b). Deliberation 2.0: Comparing the deliberative quality of online news user comments across platforms. *Journal of Broadcasting & Electronic Media*, *59*(4), 539–555. https://doi.org/10.1080/08838151.2015.1093482

Ruiz, C., Domingo, D., Micó, J. L., Díaz-Noci, J., Meso, K., & Masip, P. (2011). Public sphere 2.0? The democratic qualities of citizen debates in online newspapers. *The International Journal of Press/Politics*, *16*(4), 463–487. https://doi.org/10.1177/1940161211415849

Santana, A. D. (2014). Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Journalism Practice*, *8*(1), 18–33. https://doi.org/10.1080/17512786.2013.813194

Schudson, M. (1997). Why conversation is not the soul of democracy. *Critical Studies in Mass Communication*, *14*(4), 297–309. https://doi.org/10.1080/15295039709367020

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, *5*, 13–22.

Sobieraj, S., & Berry, J. M. (2011). From incivility to outrage: Political discourse in blogs, talk radio, and cable news. *Political Communication*, *28*(1), 19–41. https://doi.org/10.1080/10584609.2010.542360

Song, Y. Y., & Lu, Y. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

Springer, N., Engelmann, I., & Pfaffinger, C. (2015). User comments: Motives and inhibitors to write and read. *Information Communication & Society*, *18*(7), 798–815. https://doi.org/10.1080/1369118x.2014.997268

Steiner, J. A., Bächtiger, A., Spörndli, M., & Steenbergen, M. (2004). *Deliberative politics in action: Analyzing parliamentary discourse*. Cambridge University Press.

Tucker, J., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). *Social media, political polarization, and political disinformation: A review of the scientific literature. SSRN*. https://doi.org/10.2139/ssrn.3144139

Vaccari, C., Valeriani, A., Barberá, P., Jost, J. T., Nagler, J., & Tucker, J. A. (2016). Of echo chambers and contrarian clubs: Exposure to political disagreement among German and Italian users of Twitter. *Social Media + Society*, *2*(3), 1–24. https://doi.org/10.1177/2056305116664221

Van Duyn, E., & Muddiman, A. (2020). Predicting perceptions of incivility across 20 news comment sections. *Journalism*. Advance online publication. https://doi.org/10.1177/1464884920907779

Vitak, J. (2012). The impact of context collapse and privacy on social network site disclosures. *Journal of Broadcasting & Electronic Media*, *56*(4), 451–470. https://doi.org/10.1080/08838151.2012.732140

Wang, Y., & Witten, I. H. (1996). *Inducing model trees for continuous classes* (Working Paper 96/23). University of Waikato, Department of Computer Science. https://hdl.handle.net/10289/1183

Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2021). Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*, *38*, 98–139. https://doi.org/10.1080/10584609.2020.1785067

Ziegele, M., Jost, P., Bormann, M., & Heinbach, D. (2018). Journalistic counter-voices in comment sections: Patterns, determinants, and potential consequences of interactive moderation of uncivil user comments. *Studies in Communication | Media*, *7*(4), 525–554. https://doi.org/10.5771/2192-4007-2018-4-525

Author Biographies

Julia Jakob is a researcher in political communication at the University of Mannheim. Her work is particularly concerned with how the theory of deliberation can be further developed to accommodate the peculiarities of public communication in the digital age.

Timo Dobbrick is a research associate at the Mannheim Center for European Social Research. As an information scientist, his focus is in the fields of data science, machine learning and natural language processing for the automated measurement of deliberative quality.

Rainer Freudenthaler is a PhD candidate at the University of Mannheim. His research focusses on the democratic performance of online news and user-generated content using traditional quantitative content analysis, computational methods, and qualitative discourse analysis.

Patrik Haffner is a former research associate at the Mannheim Center for European Social Research.

Hartmut Wessler is a Professor for Media and Communication Studies at the University of Mannheim and a Principal Investigator at the Mannheim Center for European Social Research. A recurring theme of his research relates to the possibilities of assessing the quality of mediated contestation against diverging normative models of democracy.