Article

new media & society
1–21

DOI: 10.1177/14614448251359624
journals.sagepub.com/home/nms

Fast and furious: Temporal patterns of incivility in online comments

Ben Clarke
University of Gothenburg, Sweden

William Hedley Thompson
University of Gothenburg, Sweden

Abstract

Incivility in online comment sections is pervasive and has significant societal implications, including impacting mental well-being and increasing polarisation. This study investigates the relationship between speed of commenting and incivility, using a dataset of 38 million comments from The Guardian Online. We hypothesise that quicker responses are more likely to be uncivil and that incivility propagates through a contagion effect. Our analysis reveals that blocked comments, used as a proxy for incivility, are posted significantly faster than visible comments, for both parent and child comments. In addition, we find that the presence of blocked comments increases the likelihood of subsequent blocked comments, with decreasing time intervals between them. These findings suggest that incivility is associated with impulsive, fast thinking, while civil discourse will sometimes require slower, more deliberate practices. Our results have implications for designing online platforms to foster healthier and more productive discussions by encouraging deliberative time before readers post.

Keywords

Commenting time, deliberative democracy, digital communication, emotional reactions, incivility, incivility contagion, newsreader comments

Corresponding author:
Ben Clarke, Department of Applied Information Technology, University of Gothenburg, Forskningsgången 6, 417 56 Göteborg, Sweden.
Email: ben.clarke@ait.gu.se

Introduction

Incivility in digitally mediated interactions, such as comment sections, is a serious societal problem (Davey et al., 2023; Santana, 2014). It has acute implications such as impacting well-being (Vogels, 2022), fostering polarisation (Allison and Bussey, 2020; Chan et al., 2019) and can negatively affect democratic discussions online (Kim and Park, 2019; Kolhatkar and Taboada, 2017). Incivility's prevalence is empirically estimated at between 2% (Coe et al., 2014) to 20% of all user-generated posts (Coe et al., 2014; Kolhatkar and Taboada, 2017). In reader comments in online news, incivility's presence raises particular challenges, including affecting news organisations' credibility (Masullo et al., 2023), causing journalists to fear losing their sources (Diakopoulos and Naaman, 2011: 135–136) and increasing the presence of mis- and dis-information (Marwick and Lewis, 2017). It has even led to some news providers removing the opportunity for users to comment on their sites altogether (Bilton, 2014). Given the prevalence and seriousness of the issue, understanding how incivility manifests online is of paramount importance.

After critically reviewing theorisations of incivility in the communications literature, we draw on overlapping themes in an interdisciplinary literature to hypothesise that at least some types of incivility are likely to occur soon after some other source brings them into being. That source may be a fellow user's comment to which a second user responds uncivilly, or a journalist's article published on a news organisation's website to which a reader posts an uncivil comment. To the best of our knowledge, we are not aware of any work that has specifically proposed a direct relationship between speed of response in online user-generated spaces and an increased likelihood of incivility in the reply. After detailing our hypothesis and a review of literature helping inform it, we test its predictions on a dataset of over 38 million comments collected from The Guardian Online, which includes over 1 million comments removed by moderators. Our analysis uses variation in comment-time, by way of timestamps in the metadata of the dataset, for comments removed by The Guardian moderators (as a proxy for incivility) as compared to comments untouched by The Guardian moderators (as a proxy for civility). Furthermore, we test if incivility breeds increased incivility in our data. In short, our findings show significant support for our hypotheses and the presence of an incivility contagion effect. With this outcome, we conclude the article by speculating that media organisations and similar can use this knowledge to inform their websites' design interfaces to encourage short deliberation that will likely reduce cases of some types of uncivil contributions, encouraging the deliberative quality and productive engagement of a site's online forums for its users.

Literature review

Anti-normative communication online and theorising the place of incivility

Incivility is one type of anti-normative communication – that which does not abide by the generally held norms of considerate social etiquette because it is in some way transgressive. Such a transgression can relate to a more or less fundamental norm, leading to more or less severe anti-normative communication. Defined this way, incivility is clearly related to a number of other concepts – for example, impoliteness (Culpeper,

2015), bullying (Vogels, 2022) and toxicity (e.g. Kolhatkar and Taboada, 2017). Such behaviours are evidently not a phenomena somehow engendered with the advent of the Internet but have occurred throughout human history. However, their occurrence in a virtual context has led to an increased interest in such phenomena (Masullo Chen et al., 2019: 1; Shmargad et al., 2022: 718), given what the digital environment allows and constrains, and how such affordances differ from embodied communication.

Much of the recent scholarly literature theorises incivility online into two broad fundamental types based on the following distinction. There are serious communicative transgressions which put the principles of democracy at risk. This might be because they target users along lines of race, sex, or other protected characteristics (e.g. sexual orientation: *. . .I don't want my future son seeing 2* ***f*gs*** *walking down the street holding hands*, reported in El Sherief et al., 2018: 49) or threaten a user physically or mentally, per se (e.g. *I'll tear your limbs apart*, reported in El Sherief et al., 2018: 44) or if said user does/doesn't say/do X. In contrast are less serious communicative transgressions – for example, insults (e.g. from our data: *if you wish to ignore the science then* ***you are a wilful fool***), pointed criticisms, dismissals (e.g. from our data: *That's a stupid warmist lie [. . .] Now*, ***piss off***), and so on. While this type are rude and a nuisance, they do not sacrifice democratic values.

Unfortunately, the terminology used to refer to these two main types is often inconsistent (Hopp, 2019; Muddiman, 2017: 3182–3183; Rösner et al., 2016: 462). Papacharissi (2004), for example, refers to the former type as 'incivility' and latter type as 'impoliteness'. Oz et al. (2018) follow this distinction. However, for other scholars the approach is to reserve 'incivility' for the apparently less severe of the two aforementioned types (e.g. Mutz, 2015; Rossini, 2022), using other labels for the more severe type (e.g. for Rossini, 2022: 'intolerance'). Yet other scholars bring both types together under a two-level approach to incivility, as Muddiman (2017) argues in theorising 'personal-level incivility' and 'public-level incivility'. As Masullo Chen et al. (2019) summarise, often the question is whether 'incivility' is treated as largely synonymous with 'impoliteness' (e.g. for Mutz, 2015) or whether the latter is a concept against which to differentiate 'incivility', where incivility refers to transgressions with serious consequences for democracy (e.g. for Papacharissi, 2004). Still, as Levinson (1988) has remarked, 'terms themselves are of no analytical importance. What is essential, though, is the set of underlying discriminations' (p. 171).

More importantly, an exception must be noted to the above implied consensus concerning how the existing literature theorises the underlying discriminations. This is necessary to sufficiently capture communicative transgressions as an object of study. It concerns the target of the transgression: a specific individual or wider aspects of the communication (e.g. process, topics, tone) as a whole (Coe et al., 2014; Rösner et al., 2016: 462). The fundamental distinction explained above concerned only individuals. However, transgressions can also target wider communication, including doing so in ways that have more or less severe consequences for democracy. Muddiman's (2017) notion of 'public-level incivility', cf. 'private-level incivility', makes this apparent. Violations of democratic norms can happen 'in myriad ways' (Muddiman, 2017: 3183); not only in using hate speech that targets individuals' protected characteristics but also, for example, by blocking the right of reply, supressing discussion by changing the topic or spreading mis-information (Hopp, 2019; Muddiman, 2017).

A number of scholars undertake empirical studies in order to better inform a theoretical account of incivility. Muddiman (2017), for example, devised an experimental study to expressly bring together and test the two aforementioned common conceptualisations of incivility: disrespectful anti-normative communication (i) relating to politeness and (ii) concerning political processes and deliberative norms. Situating the study in a party-political context, Muddiman's (2017) study asked if either, or both, types of transgressions were judged by respondents as uncivil and, if both, which more so. Her results revealed that individuals saw both as notably more uncivil than communication not displaying such transgressions, and that the former type are seen as more stereotypical of incivility. In a similar study using an exploratory empirical approach to generate theoretical clarity, Hopp (2019) carried out network analysis on self-report data drawn from citizens who are active in political discussions online. Based on his sample, two clusters of incivility behaviours were determined: one concerning speech norm violations (e.g. using disrespectful language about others and using profane language) and another concerning inclusion norms (e.g. excluding discrepant others, negatively framing others' motives and discussion suppression). The study also found that transgressions intent on excluding discrepants (e.g. per se, their views) was particularly central in the variable network. Such studies therefore suggest that the public interpret a wide range of behaviours as constituting incivility, though have some awareness of a fundamental distinction between two types. Individual and group aspects can also influence the perception of incivility. For example, those who score higher on agreeableness, within the Big Five personality traits, are more inclined to judge communicative behaviours as uncivil (Kenski et al., 2020).

For the purposes of our study, we follow Papacharissi (2004) in recognising a fundamental distinction between incivility and impoliteness, and focus our empirical inquiry on the former as she defines it. We do so for two reasons motivated by our empirical study. First, as further explained below concerning our data analysis, the community under study, The Guardian Online, adopts community standards that are largely in line with Papacharissi's (2004) distinction, allowing impolite behaviour such as swearing by users but removing uncivil behaviours such as hate speech and threatening behaviour. That we adopt a theoretical definition which tightly aligns to a viable operationalisation of the phenomenon under study – in the form of comments removed by The Guardian's moderators, in line with their community standards – is important for a second reason: doing so allows us to take a large real-world dataset of incivility classifications. This approach gives us approximately 1 million comments classified as uncivil which in turn means that any discrepancies in the application of moderating standards will be negligible.

Thinking fast but computing slower

Kahneman (2011) proposed that many aspects of human psychology, including judgement and decision-making, are subject to more instinctive or more deliberate cognitive processing, his now well-known distinction between System-1 (instinctive; thinking fast) and System-2 (deliberate; thinking slow) processes. Kahneman (2011: 19–20) gives two canonical examples: reacting to strongly valanced emotional displays on human

faces (System-1) and solving a complex mathematical problem (System-2). Kahneman stresses that the two systems are interdependent such that one's attention and effort is shared across them; System-1 may often be in control but pass responsibility to System-2 when it runs into problems – for example (Kahneman, 2011: 22), if one hears a loud, offensive comment and instinctively goes to orient their attention to it (System-1) but, precisely because it is offensive, overrides this instinct (System-2). Similarly, it is difficult for one's System-1 to be attendant when System-2 is heavily engaged in a task, given the demands put on attention by System-2 – for example, if one is engaged in a routine conversation with their friend while watching the big sports game, a pause in the former is likely when observing the crucial (e.g. goal-scoring) moment in the latter. *Thinking slow* does not necessarily require more literal time; let us call this absolute-time, in the Newtonian sense. It is, however, more effortful, more attention-demanding and more deliberate and therefore typically involves more absolute-time. The same is true in reverse: *thinking fast* does not necessarily entail shorter absolute-time, though often does, as System-1 judgements are typically more automatic, instinctive reactions.

Kahneman (2011) provides examples of a number of facets of the distinction that relate explicitly to language use: understanding a written word in front of us in our mother tongue, completing phraseological expressions like 'rise and.'.., comprehending simple sentences and holding routine conversation are all aspects typically handled by System-1. In contrast, System-2 is required for following the voice of a particular person in a crowded and noisy room, for identifying the number of occurrences of a letter or word on a page; for checking the validity of a complex logical argument. Relating to incivility, Kahneman (2011) remarks that it is *thinking slow* which is 'credited with the continuous monitoring of [one's] own behaviour–the control that keeps you polite when you are angry' (p. 24). Indeed, he gives the labour of control required to not blurt out an offensive remark when we are annoyed as a hallmark example of deploying System-2 (Kahneman, 2011: 25). In the context of reader commenting sections, Zhang et al. (2023) used Kahneman's ideas on dual processing to investigate moderators' abilities to identify problematic transgressive comments when working under time pressure. They devised a system, *BiasX*, to slow moderators' cognitive thinking and force them to rationalise their arguments for classifying a comment as either problematic or unproblematic. Doing so improved moderators' ability to disentangle: (i) genuinely problematic comments even if not apparent as such on the surface and (ii) fundamentally unproblematic comments even if initially appearing to be problematic.

Summarising Kahneman's (2011) distinction as relevant here, a number of aspects of productive discussions, including being civil (pp. 24–25), cost. They cost, namely, human effort; they are mentally taxing, requiring the deliberate action of humans to deliberate, hallmarks of System-2 thinking. This is counter to frequent human inclination towards more instinctive, prime facie easier, action – of the kind 'of least effort' to the self. These could be indulging one's frustrations at being challenged in argument or with views one finds unfavourable and thus providing uncivil responses, or in exaggerating the arguments of others' so as to dismiss them rather than trying to contemplate them for their potential merits. These latter behaviours are hallmarks of System-1 thinking. It should be noted that, while influential, dual processing theories have undergone theoretical developments since Kahneman (2011) and faced critical scrutiny (e.g. Grayot, 2020).

Critics often advocate for more spectrum-based models of cognitive processing, rather than strict binary categories (e.g. Melnikoff and Bargh, 2018). At the same time, developments in cognitive science have seen dual-process distinctions integrated into broader frameworks, such as predictive coding and active inference models (Tschantz et al., 2023). Regardless of the precise theoretical stance, it is widely acknowledged that some cognitive processes are faster and more automatic, while others are slower and more deliberative. Kahneman's (2011) work usefully illustrates this general distinction.

From the perspective of individual psychology more generally, Frischlich et al. (2021) and Park and Martinez (2022) found that, for users instigating incivility online, there was a correlation with lower psychological well-being, with dark personality traits, and feelings of victimisation. These studies also found incivility tendencies were skewed in favour of being male and younger. Psychological processes such as self-regulation and emotional regulation have been found to correlate with civil behaviour, and a reduction in the effective functioning of such processes with uncivil behaviour (Meier and Gross, 2015). Similarly and in a face-to-face classroom context, Spadafora et al. (2018) found that less emotional and behavioural regulation correlated with incivility. Furthermore, Tice et al. (2001) found that impulsivity in reactions can increase when trying to regulate negative emotions. For Kahneman (2011: 40–44), the successful employment of such self-regulation processes is linked to System-2 work, and these help one avoid slips like 'using sexist language' (p. 41). Taken together, our hypothesis is informed by an expectation that uncivil actions are made quickly and impulsively, with little to no reflection or emotional control, and that, while all users will find themselves in situations that make these things more likely, users with certain personality characteristics are even more prone to such behaviour.

In relation to human-computer interaction, Hallnäs and Redström (2001) challenged the assumption that technological design should only be aimed at efficiency gains, arguing its status as default design strategy for technology resides in the origins of computing: office-work and science. They argued that other strategic uses of technology, such as artistic appreciation, should also be catered for by designers focusing on different strategic goals. Particularly, they promoted a design programme for 'slow technology', technology designed to facilitate users' reflection and mental rest, which they argued was critical in an age of ubiquitous computing and 'in a more and more rapidly changing environment' (Hallnäs and Redström, 2001: 202). The aim is not to reduce cognitive load but, rather, to encourage users to employ suitable quantities and qualities of time for optimal fulfilment of the task in question given its cognitive demand. Such a design strategy mostly runs counter to technology designed according to efficiency: an extended learning time rather than a shortened one; reflection rather than production; time *giving* – in the sense of opening up for the presence of time – rather than saving time; expanding time rather than compressing time (Hallnäs and Redström, 2001: 203).

In the recent context of further technological ubiquity and more advanced networked computation, Kitchin and Fraser (2020) develop these arguments, including speculating that digital technologies can cause us to be in several moments at once, bringing about cognitive dissonance (Kitchin and Fraser, 2020: 39). Such networked technologies become increasingly omnipresent and fundamental, both for the individual's daily life (e.g. informing a friend of late arrival for an arranged coffee) and the functioning of

society (e.g. organising transport systems). The consequent trend is towards exponential *acceleration*. Acceleration has positives associated with, for example, overcoming geographical boundaries (Kitchin and Fraser, 2020: 33) and in quickly accessing more detailed information to complete a range of tasks (Kitchin and Fraser, 2020: 35). However, there are notable downsides, including an expectation that we respond more quickly and in unqualified terms; that is, without discriminating with respect to ours or others' being in public vis-a-vis private domains; to whether or not one already has currently on-going tasks; to what the task requiring a response demands of one's attention and efforts; and so on. This emphasis on speed of response causes an expanding, self-perpetuating response loop, emphasising the constant present. In practice, one's attention frequently becomes split, their efforts spread too thinly across more and temporally competing tasks which itself can become demoralising. There is limited time for detachment, unwinding and personal reflection. Ultimately, as Kitchin and Fraser (2020) conclude, the potential effects for the individual are significant: one is often under increased pressure and anxiety as a result of being rushed and harried, leading to exhaustion or worse (p. 37–42). In terms of likely effects for communication, that interlocutors communicate in less emotionally favourable states is well-known to increase the likelihood of negative consequences for the subsequent communication (Wood, 2014). In addition, in such a digital communication landscape 'there is no time for reflection, contemplation, slow rational deliberation [or] considered answers' (Kitchin and Fraser, 2020: 40). Where communication is assisted by automation (e.g. algorithms) the problems are multiplied as such automated practices are usually rule-based, de-contextualised and therefore further disenable reflection, deliberation and communal debate (Kitchin and Fraser, 2020: 40–41).

Bringing together the above prime facie disparate strands of research, we predict that a longer absolute-time is an all-but necessary condition for true constructive discussions in asynchronous digital contexts such as newsreader comments forums. Without this, there is a significantly increased risk of less constructive contributions. One sub-set of such comments are the uncivil kind, which by definition compromise the deliberative qualities of social conversation. Thus:

> ***Hypothesis 1:*** uncivil communication is more likely to occur in reader comments soon after the source to which they respond and, conversely, uncivil communication is increasingly unlikely in reader comments the greater the time duration between source and reply.

As a related aspect, several studies argue that transgressive communication such as incivility has a self-perpetuating characteristic (Foulk et al., 2016; Kim and Park, 2019; Park and Martinez, 2022; Rosen et al., 2016). Specifically, the presence of an initial communicative transgressive tends to attract more of the same in return, increasing the stakes, emotions and impulsivity involved. Ekman (2003) argues that this is a characteristic, which typifies anger as an emotion (p. 111). This pattern has been observed in a range of user-generated digital genres, including newsreader comment forums (e.g. Gervais, 2015; Masullo Chen and Lu, 2017). The latter study, for example, conducted an experiment on reader comments in relation to abortion, finding uncivil disagreement, but not

civil disagreement, led participants to behave uncivilly in reply. Therefore, we additionally hypothesise the following with the specification that temporality plays a part in the contagion (e.g. response times may get increasingly quicker in uncivil posts):

> ***Hypothesis 2:*** uncivil communication increases the likelihood of further uncivil communication.

Method

In this section, we first describe the salient characteristics of the dataset used for testing our hypotheses and the procedures adopted to collect it. Second, we explain our analytical procedures, including how we operationalised both incivility and time, as well as a description of the statistical procedures carried out to generate our findings.

Data collection: *the Guardian Opinion* comment corpus

We utilised a dataset of online newsreader comments posted to The Guardian's webpages. The Guardian Online had 580.4 million visits globally in January 2024 (Semrush, 2024). In July 2021, The Guardian sold 105,000 printed copies daily in the United Kingdom (Press Gazette, 2023). Its readers are considered left-leaning, politically; in post-World War 2 U.K. general elections, the newspaper has only twice (1951 and 1955) not endorsed parties of the centre or centre-left (The Guardian, 2010). A 2018 Ipsos-MORI poll surveying online users about trust in digital news platforms judged The Guardian as having the most reliable content. Its website went live in January 1999, first allowing readers to comment on online versions of its articles in March 2006, dubbing the feature *Comment is Free* (The Guardian, 2006). By April 2016, The Guardian's site received approximately 70 million comments (Mansfield, 2016). As discussed more below, The Guardian has been reflective in how comments can be posted to its site, and how these are to be organised, arranged and particularly how they are to be moderated. Arguably, this is in line with its political leanings; Diakopoulos (2015) discusses the same concerning what might be considered its U.S. equivalent, The New York Times (p. 1154). Moderation-wise, a number of changes followed Gardiner et al.'s (2016) study commissioned by The Guardian into reader comments on its articles. One finding was that, of The Guardian's 10 most abused journalists by commenters, eight were women and the two men were Black, while the 10 least abused journalists were all White men. The study also found that, on average, 2% of comments were blocked by its moderators.

To test our hypotheses, we downloaded all reader comments from all articles between 1st March 2006 to 18th March 2024 from the Opinion section of The Guardian's homepage (http://theguardian.com/commentisfree). The resulting corpus contains 38 million comments, of 2.4 billion words. Comments were downloaded through The Guardian's API, for which they kindly provided us an API key. Table 1 provides descriptive statistics about the number of comments by type. In addition to the textual comments, the dataset contains timestamps for comments, usernames of posters, upvoting on comments and information on threadedness. Since 2009, The Guardian (2009b) introduced time-windows within

Table 1. Description of the corpus analysed here, obtained from the Opinion section at The Guardian's website.

Number of articles	130,974
Number of comments	38,839,387
....of which are flat comments in unnested sections (2005–2012)	5,987,184
.....of which are blocked	171,900
....of which are parents in nested comment sections (2012–2024)	10,733,678
.....of which are blocked	477,028
.....of which contain children	5,200,855
....of which are children in nested comment sections (2012–2024)	22,118,515
.....of which are blocked	358,385
Total number of blocked comments	1,007,313
Number of users	713,146
...of which have at least one blocked comment	156,738

which reader comments must be received after the article publication time. We therefore excluded all comments over a week old and removed other anomalies found in the dataset. In 2012, The Guardian made it possible to thread comments, to a single-level of nesting. Our dataset therefore contains comments to an article ('parents'; since 2012) and also comments that are replies to comments ('children'; since 2012) with this nested format implemented since 2012. Before 2012, there was no possibility to answer specific comments (i.e. 'flat'). We therefore split our analyses of comments into three different types: children, parents and flat.

Data analysis: moderated comments as a proxy for incivility

In our dataset, a number of comments (=1,007,313) were removed by The Guardian moderators. The Guardian (2009a) policy for users commenting on their news articles claims that only comments containing any of the following will be removed from their webpages: hate speech (e.g. racism, sexism and homophobia), spamming, material from chatbots, off-topic comments, personal attacks, trolling and threatening behaviour. As discussed in our literature review above, such an operationalisation is consistent with the definition of incivility adopted here, following Papacharissi (2004), in that all these types – aside from one (see just below) – are considered instances of incivility. However, the reverse is not true: that is, The Guardian's moderation standards do not account for all communicative behaviours that the literature considers to be cases of incivility. On the flip-side, therefore, instances of profanity, swearing and otherwise difficult to categorise insulting behaviour – impoliteness rather than incivility, following Papacharissi (2004) – are likely to remain on the site because they do not expressly break The Guardian's (2009b) community standards (e.g. 'on the whole we don't remove comments purely on the basis of swearing where this is part of the cut and thrust of conversational debate'). While our hypothesis, concerned with incivility as a communicative transgression, accounts for many of the aforementioned types of comment that The Guardian claims to remove, spamming and bot-activity are an apparent exception. Their presence could

therefore confound our analyses. We therefore performed additional checks to control for potential spam or commercial activity: first, we filtered users by the percentage of comments that they had which were visible. Second, we filtered users by the total number of recommendations they had received for all their comments. The logic of these methods was that users with either, or both, a higher proportion of visible comments and/or high number of recommendations are likely to be reputable and therefore not likely to reflect bot, spamming, or commercial activity. These filtering steps were only applied in the sub-analyses to check for bot, spamming or commercial activity (see Figure 1(c), (d), (g), (h), (k) and (l)). For all other analyses no filtering was performed.

To analyse time in a methodical way, we distinguished parent comments from child comments. Response-time for parent comments was calculated by comparing the posting time of the comment under analysis with the time of the publication of the article on The Guardian's website; in contrast, response-time for child comments was calculated by using the posting time of the child comment in question relative to the time that the parent comment it replies to was posted. A few anomalies remained: first and regarding time to publication, in some cases the article publication time was later than the first comment, which reflected when an article had been updated and not the original published time. In these cases, we took the first comment time as a proxy for the article publication time. Second and regarding time to nested parent comment, there were 10 instances (one of which was blocked), which we manually reviewed, where the child-comment was a short time before the parent. Since these cannot be explained nor easily corrected, they were removed from the dataset.

To calculate incivility contagion, we restrained the analysis to the first nine blocked comments in a thread because taking more blocked comments than this left us with fewer than 100 threads for each step. Specifically, there were 62 comment threads that had between 10 and 29 blocked comments. This became too noisy. For this analysis, we calculated the conditional probability of at least n comments being blocked given that there were at least n-1 comments which were blocked. Thereafter, we analysed the median time between the nth blocked comment in relation to the n-1th blocked comment (i.e. the time between the ninth and the eighth comment, eighth and seventh). For the first blocked comment, we contrasted that with the posting time of the parent comment. This analysis was restrained to the child comments because, for parent and flat comments, it is less clear if there is a direct relation between subsequent comments or if there are different conversations on-going in the comment section (see Marcoccia, 2004 on polylogues).

In terms of our statistical calculations, non-parametric tests were chosen throughout due to the distributions in the data. For comparisons between two groups, we used the Mann–Whitney test (in Figure 1(a), (e) and (i)) with rank bi-serial correlations to estimate effect size. For comparisons between the three groups, we applied the Kruskal–Wallis test with Dunn's post hoc test for individual comparisons (in Figure 2(a), (c) and (e)). When comparing individual positive-negative differences, the Wilcoxon signed-rank test was used (in Figure 2(b), (d) and (f)). For the incivility contagion analysis, we used Spearman Rank to calculate whether the median times were decreasing as a function of n (in Figure 3(c)). For bootstrapping the differences in the densities (Figure 1(b), (f) and (g)), we randomised group membership (blocked vs visible) while maintaining the group sizes for 1000 permutations. To create confidence intervals while accounting

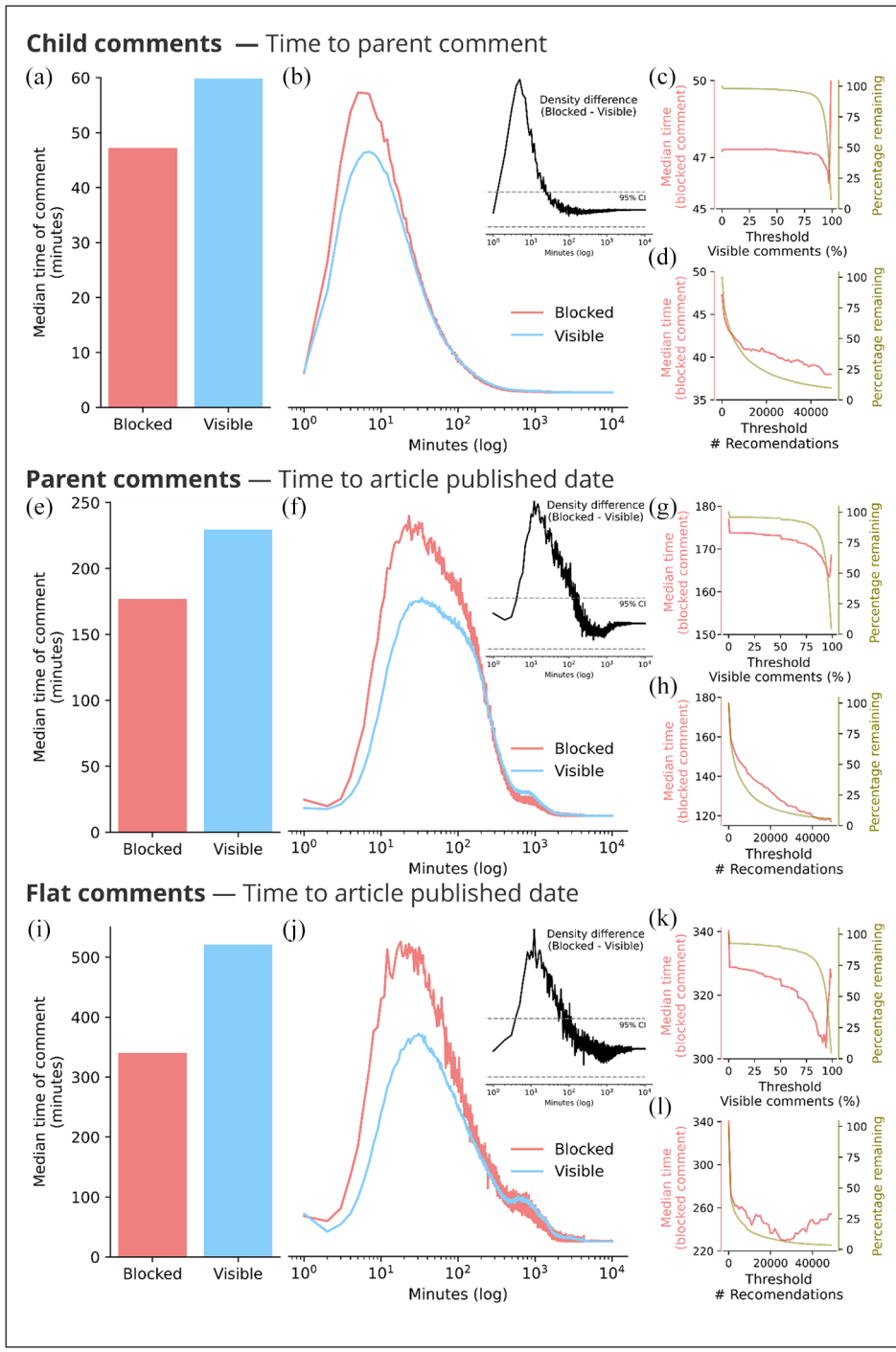

Figure 1. *(Continued)*

Figure 1. Difference between blocked and visible comments for (a) median posting time between parent comment and child, (b) The density of comment response times between parent and child comment in nested comment sections along a logged x-axis in minutes (filtered at 1 week for both groups). The inlet shows the difference between the two distributions and the 95% confidence interval of the maximum difference in densities per permutation when bootstrapping comment membership to groups of each sizes to the blocked and visible comments. (c) (d) (e) The same as A, but for flat comments, the response time is between the publication date and comment time. (f) The same as B, but for flat comments, the response time is between the publication date and comment time. (g) (h) (i) The same as A, but for parent comments, the response time is between the publication date and comment time. (j) The same as B, but for parent comments in threads, the response time is between the publication date and comment time. The resolution of all density figures is per minute (k) (l).

for multiple comparisons, we took the maximum and minimum differences between the densities in each permutation and calculated the 97.5th and 2.5th percentile values, respectively.

Results

This section has three main parts. The first two offer the outcome of analyses to test our first hypothesis, concerning the speed of posting comments by readers on The Guardian's website and the likelihood of incivility. The third part of our results corresponds to our second hypothesis, the issue of an incivility contagion phenomenon.

Uncivil comments are quicker: blocked comments are posted quicker than visible comments (H1)

We observed a statistically significant difference in the time to comment visibility between the two groups, with blocked comments being quicker. For child comments, the median time for blocked comments was 47 minutes and 15 seconds ($N=358{,}385$), while for visible child comments, it was 59 minutes and 52 seconds ($N=21{,}760{,}130$; Figure 1(a) $U=3.57 \times 10^{12}$; $p<.001$). The effect size ($r_{rb}=0.083$) indicates a small to medium effect. In addition, subtracting the densities from each other reveals that the time span significantly differs (Figure 1(b)), showing that blocked comments were more likely to occur between 2 and 24 minutes after the parent comment ($p<.001$ corrected, 95% CI=[−0.00058, 0.00060]).

As raised above, there is a possibility that automatic bots post quickly and this spamming activity is blocked by moderators, which could explain the decrease in times. To account for this, we chose two variables, the total number of recommendations a user had and the ratio of positive comments a user had, recalculating the median time difference. When applying a threshold to try to remove possible bot activity (i.e. users with low recommendations and few visible comments) this leads to a decrease in the times, which demonstrates that the quick blocked comments appear to be the uncivil and inappropriate comments (Figure 1(c) and (d)).

Similar patterns were observed for both parent comments and flat comments. For parent comments, the median difference was larger than for child comments: the median

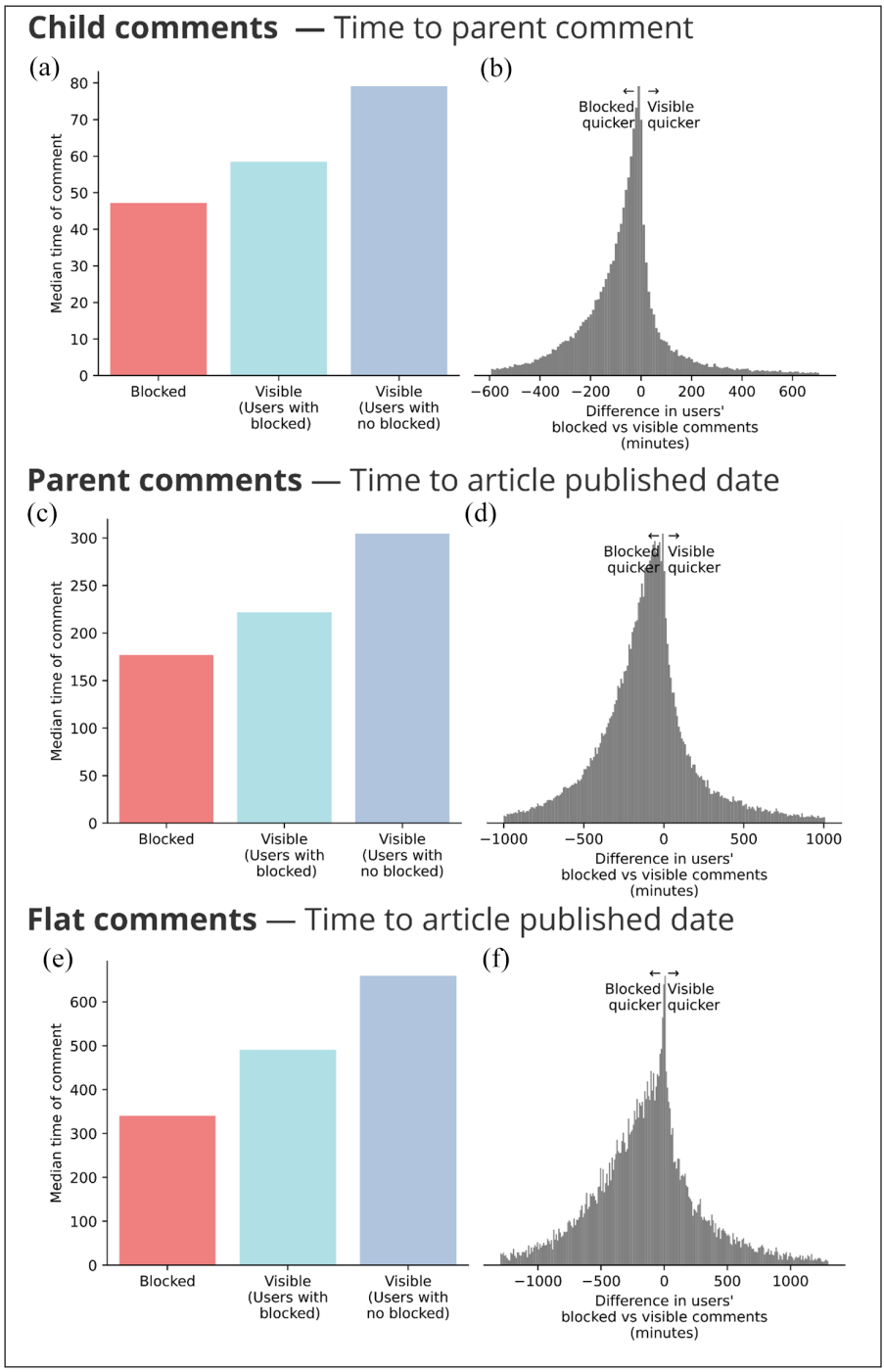

Figure 2. *(Continued)*

Figure 2. (a) Median timings for visible comments of child posts when splitting visible comments into users with blocked comments and users with no blocked comments. (b) Density plot of users displaying the difference in average users' blocked minus visible child comments. Density plot has been scaled to only show 95% of data to remove more extreme values. A negative score implies that the user wrote blocked comments quicker than their visible comments. (c, d). The same as A, but for parents in nested comments and using time to article publishing date. (e, f). The same as CD but for flat comments.

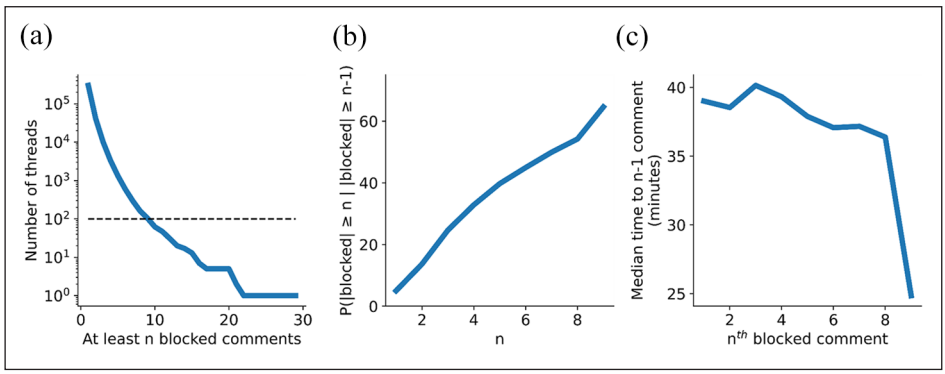

Figure 3. (a) The number of threads (i.e. children connected to a parent) where there are at least *n* number of blocked comments. This decreases logarithmically. The dotted line indicates where there were at least 100 threads with *n* blocked comments ($n=9$). (b) Conditional probability showing the likelihood of at least *n* blocked comments, given that there were at least *n-1* blocked comments. We have restrained to show only comments up to the first nine blocked comments (at least 100 threads). (c) Median time of the n^{th} blocked comment in a thread compared to the $n\text{-}1^{th}$ blocked comment. The time to the first blocked comment is to the parent comment. Again, we are restricting to the first nine blocked comments (at least 100 threads).

comment time for blocked parents ($N=477{,}028$) was 176 minutes and 58 seconds after the article publication date and 229 minutes and 37 seconds after the article for visible comments ($N=10{,}256{,}650$; $U=2.13 \times 10^{12}$; $p<.001$; Effect size (r_{rb})$=0.129$). In the density distributions, a significant period of 5 to 118 minutes following the article showed an increased propensity for blocked comments ($p<.001$ corrected, CI$=[-0.000268, 0.000270]$). This significant period lasted for a total of 140 minutes, until 183 minutes. A similar pattern was observed in the flat comments sections (Median blocked: 340 minutes and 15 seconds; Median visible: 521 minutes and 9 seconds; MW$=4.38 \times 10^{11}$, $p<.001$; Blocked: $N=171{,}900$; Visible: $N=5{,}815{,}284$; Effect size (r_{rb})$=0.123$). A continual significant period between the densities was observed between 4 and 52 minutes, (74 significant time points in total lasting to 113 minutes; $p<.001$ corrected; 95% CI$=[-0.000366, 0.000402]$). Finally, to control for possible bot or commercial activity, for both parent and flat comments, a similar picture to child comments emerges when a threshold based on recommendations or proportion of visible comments is applied (Figure 1(j), (h), (k) and (l)). In sum, support for Hypothesis 1 is found.

Blocked comments are quicker compared to a user's visible comments (H1)

One question that emerges in this data is whether this shift in the distributions for blocked comments is indicative of a profile of user who is generally responding quicker, or whether this actually reflects an affective state in a user that increases the propensity for a quick response. To investigate this, we split the visible comments into those from users with at least one blocked comment and from users with no blocked comments. As stated above, most reader comments on The Guardian are made by users with at least one blocked comment.

When splitting the visible comments by user type, the only difference that emerges is that users with no blocked comments tend to have even slower times (see Figure 2(a), KW = 55,926 $p < .001$). Dunn's post hoc test showed comparisons between all categories (blocked, visible comments for users with one blocked comment, visible comments for users with no blocked comments) as significant ($p < 0.001$, Bonferroni corrected). We then analysed the difference between visible and blocked comments for each user with one or more blocked comment, which revealed that 73.2% had significantly average slower comment times for visible comments (Wilcoxon $W = 5.37 \times 10^8$, $p < .001$; see Figure 2(b)). A similar pattern emerged for parent comments (Figure 2(c) and (d)). For median comment times, there was a significant difference between all three categories (KW = 89,340, $p < .001$, with Dunn's post hoc $p < .001$, Bonferroni corrected). For the difference in user comments, 70.6% were quicker for blocked comments (Wilcoxon $W = 1.13 \times 10^9$, $p < .001$).

Finally, for flat comments, the pattern was also preserved (Figure 2(e) and (f)). For the median response time, again there was a significant difference between all three categories (KW = 41,555 $p < .001$), with Dunn's post hoc test ($p < .001$, Bonferroni corrected). Furthermore, 65.6% of users had quicker response times for blocked comments (Wilcoxon $W = 1.20 \times 10^8$, $p < .001$). In sum, further support for Hypothesis 1 is found, including that it is a phenomenon that is true for a wider profile of users.

Incivility contagion: blocked comments increase the likelihood of more blocked comments with decreasing time intervals (H2)

When a parent comment has at least one child that is blocked, it increases the chance that more blocked child comments will follow in the chain (Figure 3(a) and (b)). The probability that a parent with any child comments has at least one blocked child comment is 5.00%. If there is at least one blocked child, the probability of at least two blocked child comments increases to 13.66% and so on (Figure 3(b)). When there are over six blocked comments in a thread, then it is almost 50% likely there will be more blocked comments (49.92%) and stays above 50% until the cut-off point when at nine blocked comments (Figure 3(b)).

Finally, the question arises whether the increased likelihood of blocked comments in threads reduces the timespan between the replies. Here, we calculated the median time of the n^{th} comment in relation to the $n\text{-}1^{th}$ comment, instead of to the parent comment (Figure 3(c)). Here we see that these time differences shorten significantly between subsequent blocked comments (Spearman $\rho = -0.85$, $p = .004$). Together, these results show

that when there are blocked uncivil comments, the chances increase that there will be more uncivil comments and that the time difference between these comments will shorten. In short, incivility brings about more and quicker incivility. Support for Hypothesis 2 is therefore found.

Discussion and conclusion

Our findings provide convincing support for our hypothesis that quicker responses in interactive newsreader comment sections are more likely to contain instances of incivility: the median comment time of all blocked comments is statistically significantly quicker than the median comment time of all visible comments in the community. We have also demonstrated that these results are not due to bot or commercial activity. For users who have both blocked and visible comments, the average comment time of their blocked comments is statistically significantly quicker than the average comment time of their visible comments. Here, we saw that, at the individual level, despite the possibility of unknown time for when a user will read a specific article to comment, there is still an increased likelihood of quicker blocked comments than visible comments. Finally, we see evidence of an incivility contagion, where the occurrence of blocked comments increases the likelihood of subsequent blocked comments, and does so with decreasing time intervals.

In considering the challenges posed to user-generated content online by the presence and frequency of incivility, questions concerning free-speech and the extent and forms of moderation naturally arise. Seering et al. (2019) broadly distinguish two approaches to online moderation: reactive interventions (i.e. delete all the bad stuff and ban those who post it) and proactive interventions (i.e. determine any means for stimulating healthier, more productive discussions). Studies suggest that promising proactive interventions include: highlighting user comments judged by moderators as engaging in civil deliberative discussion (Diakopoulos, 2015), developing classifiers to promote reader comments displaying constructive qualities (i.e. on-topic, specific and based on personal experience and/or knowledge; Kolhatkar et al., 2023), having journalists engage with readers on comments sections (Ksiazek, 2018) and presenting CAPTCHAs that prime positive emotions before users can comment (Seering et al., 2019). The last of these studies, which force users to stop and enter a CAPTCHA, may have had the added benefit, not in focus in the researchers' study design, of slowing users down, which thus may have helped enable the more deliberative processes that they find. Another perspective is to emphasise moderators' practice. An example of this is seen in the promising results in Zhang et al'.s (2023) work, discussed above in our literature review, theirs being the closest proposals to our own of which we are aware. Indeed, communication is always at least a two-way process, and the aforementioned study highlights that what is demanded of a particular type of receiver role in constructive deliberation can also be negatively impacted under short time constraints.

Several shortcomings of our study should be noted: first, although we know when users posted comments, we do not know when they read the article or comment to which they respond; no such 'reading-receipt' data for user comments exists. However, we

cannot conceive of how our observed effect could be induced due to this time lag. Rather, it could well be that the relationship we hypothesise here is notably stronger than we can here show. Second and related, our analysis does not account for differences in writing time for comments; that is, potentially blocked comments may in part be quicker because they are on average shorter. Without either (i) timestamp data for when a user starts to write a comment or even (ii) access to the word- or character-lengths of blocked comments, we cannot compute exact or approximate writing times to factor this into our analysis. Third, of The Guardian's stated breaches to its community standards (e.g. off-topic, hate speech, threatening; The Guardian, 2009a), we do not know which commonly occurred in the case of blocked comments, although our analyses guard against any moderate chance that blocked comments are explained by bots posting spam as comments. Related to this and because we cannot access the textual data of blocked comments, fourth, we cannot be sure that The Guardian moderators carried out their work consistently or in strict accordance with their professed standards. Fifth, we cannot be sure exactly how The Guardian carries out their moderation practices (e.g. with automatic tools, all or somewhat by hand), nor whether their processes or moderation standards have changed across time. The Guardian (personal communication) have relayed that this data is too sensitive for them to share with us. Finally, not all types of incivility as these have been theorised in the existing literature are accounted for by The Guardian's (2009a) community standards, meaning that our analysis is not a comprehensive account of all incivility dimensions.

Many relevant future studies can be envisaged. For example, do similar phenomena exist on different user-generated platforms, such as video platforms like *TikTok*? Our results motivate lab-setting behavioural experiments which could control for reader-receipt as well as monitor reactions and text production while testing time-delay interventions. Furthermore, media organisations can harness the knowledge from this study and test comment interfaces that encourage deliberative time. These could be simple proposals where commenters experience a short time-delay before posting or, similar to Seering et al. (2019), where users are given a short engagement task to achieve the same end. Our findings here suggest that doing such things will likely lead to reduced cases of some types of uncivil contributions, and so better provide deliberative democratic spaces aimed at genuinely promoting productive and sustained user engagement.

Author contributions

The majority of the literature review and the generation of the hypothesis was devised by Clarke, with some support by Thompson. Data collection and analysis was primarily carried out by Thompson, with some support by Clarke. Both authors prepared the manuscript jointly.

Data availability statement

The analysis code and code to obtain the data can be found at https://github.com/wiheto/guardian_scrape/ [Note that the GitHub address is not live at the moment during the peer review]. Downloading the data will require an API key. We cannot share the data openly for copyright reasons.

Declaration of conflicting interest

The author(s) declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval statement

This article does not contain any studies with human or animal participants. Rather, the research was conducted on readily available online data, for which we judged no ethical approval needed to be sought in line with academic discussions in the field[1].

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Ben Clarke https://orcid.org/0000-0003-0992-5598

William Hedley Thompson https://orcid.org/0000-0002-0533-6035

Note

1. British Academy of Applied Linguistics. (2021). Recommendations on good practice in applied linguistics. https://www.baal.org.uk/wp-content/uploads/2021/03/BAAL-Good-Practice-Guidelines-2021.pdf

References

Allison K and Bussey K (2020) Communal quirks and circlejerks: a taxonomy of processes contributing to insularity in online communities. *Proceedings of the International AAAI Conference on Web and Social Media* 14: 12–23.

Bilton R (2014) Why some publishers are killing their comment sections. Available at: https://digiday.com/media/comments-sections/

Chan C, Chow CS and Fu K (2019) Echoslamming: how incivility interacts with cyberbalkanization on the social media in Hong Kong. *Asian Journal of Communication* 29(4): 307–327.

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

Culpeper J (2015) Impoliteness strategies. In: Capone A and Mey JL (eds) *Interdisciplinary Studies in Pragmatics, Culture and Society*. London: Springer, pp. 421–445.

Davey J, Miller C and Guhl J (2023) Hate of the nation: a landscape mapping of observable, plausibly hateful speech on social media. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0029/268256/Hate-of-the-Nation.pdf

Diakopoulos N (2015) The editor's eye: curation and comment relevance on the New York Times. In: *Proceedings of the 18th ACM conference on computer supported cooperative work and social computing*, Vancouver, BC, Canada, 14–18 March, pp. 1153–1157. New York: ACM.

Diakopoulos N and Naaman M (2011) Towards quality discourse in online news comments. In: *Proceedings of the ACM 2011 conference on computer supported cooperative work*, Hangzhou, China, 19–23 March, pp. 133–142. New York: ACM.

Ekman P (2003) *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. New York: Times Books/Henry Holt.

El Sherief M, Kulkarni V, Nguyen D, et al. (2018) Hate lingo: a target-based linguistic analysis of hate speech in social media. arxiv.org. Available at: https://ojs.aaai.org/index.php/ICWSM/article/view/15041/14891

Foulk T, Woolum A and Erez A (2016) Catching rudeness is like catching a cold: the contagion effects of low-intensity negative behaviors. *Journal of Applied Psychology* 101(1): 50–67.

Frischlich L, Schatto-Eckrodt T, Boberg S, et al. (2021) Roots of incivility: how personality, media use, and online experiences shape uncivil participation. *Media Communication* 9(1): 195–208.

Gardiner B, Mansfield M, Anderson I, et al. (2016) The dark side of Guardian comments. Available at: https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments

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

Grayot JD (2020) Dual process theories in behavioral economics and neuroeconomics: a critical review. *Review of Philosophy and Psychology* 11(1): 105–136.

Hallnäs L and Redström J (2001) Slow technology–designing for reflection. *Personal and Ubiquitous Computing* 5: 201–212.

Hopp T (2019) A network analysis of political incivility dimensions. *Communication and the Public* 4(3): 204–223.

Kahneman D (2011) *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kenski K, Coe K and Rains SA (2020) Perceptions of uncivil discourse online: an examination of types and predictors. *Communication Research* 47(6): 795–814.

Kim JW and Park S (2019) How perceptions of incivility and social endorsement in online comments (dis) encourage engagements. *Behavior and Information Technology* 38(3): 217–229.

Kitchin R and Fraser A (2020) *Slow Computing: Why We Need Balanced Digital Lives*. Bristol: Bristol University Press.

Kolhatkar V and Taboada M (2017) Constructive language in news comments. In: *Proceedings of the first workshop on abusive language*, pp. 11–17. Available at: https://aclanthology.org/W17-3002/

Kolhatkar V, Thain N, Sorensen J, et al. (2023) Classifying constructive comments. *First Monday*. Available at: https://firstmonday.org/ojs/index.php/fm/article/view/13163

Ksiazek TB (2018) Commenting on the news: explaining the degree and quality of user comments on news websites. *Journalism Studies* 19(5): 650–673.

Levinson S (1988) Putting linguistics on a proper footing. In: Drew P and Wotton A (eds) *Erving Goffman: Exploring the Interactive Order*. Cambridge: Polity Press, pp. 161–227.

Mansfield M (2016) How we analysed 70m comments on the Guardian website. Available at: https://www.theguardian.com/technology/2016/apr/12/how-we-analysed-70m-comments-guardian-website

Marcoccia M (2004) On-line polylogues: conversation structure and participation framework in internet newsgroups. *Journal of Pragmatics* 36(1): 115–145.

Marwick A and Lewis R (2017) *Media Manipulation and Disinformation Online*. New York: Data & Society Research Institute.

Masullo Chen G and 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.

Masullo Chen G, Muddiman A, Wilner T, et al. (2019) We should not get rid of incivility online. *Social Media + Society* 5(3): 862641.

Masullo GM, Tenenboim O and Lu S (2023) 'Toxic atmosphere effect': uncivil online comments cue negative audience perceptions of news outlet credibility. *Journalism* 24(1): 101–119.

Meier LL and Gross S (2015) Episodes of incivility between subordinates and supervisors: examining the role of self-control and time with an interaction-record diary study. *Journal of Organizational Behavior* 36(8): 1096–1113.

Melnikoff DE and Bargh JA (2018) The mythical number two. *Trends in Cognitive Sciences* 22(4): 280–293.

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

Mutz DC (2015) *In-Your-Face Politics: The Consequences of Uncivil Media*. Princeton, NJ: Princeton University Press.

Oz M, Zheng P and Chen M (2018) Twitter versus Facebook: comparing incivility, impoliteness, and deliberative attributes. *New Media & Society* 20(9): 3400–3419.

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

Park LS and Martinez LR (2022) An 'I' for an 'I': a systematic review and meta-analysis of instigated and reciprocal incivility. *Journal of Occupational Health Psychology* 27(1): 7–21.

Press Gazette (2023) Newspaper ABCs. Available at: https://pressgazette.co.uk/media-audience-and-business-data/media_metrics/most-popular-newspapers-uk-abc-monthly-circulation-figures-2/

Rosen CC, Koopman J, Gabriel AS, et al. (2016) Who strikes back? A daily investigation of when and why incivility begets incivility. *Journal of Applied Psychology* 101(11): 1620–1634.

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

Rossini P (2022) Beyond incivility: understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research* 49(3): 399–425.

Santana AD (2014) Virtuous or vitriolic. *Journalism Practice* 8(1): 18–33.

Seering J, Fang T, Damasco L, et al. (2019) Designing user interface elements to improve the quality and civility of discourse in online commenting behaviors. In: *Proceedings of the 2019 CHI conference on human factors in computing systems*, Glasgow, 4–9 May, pp. 1–14. New York: ACM.

Semrush (2024) theguardian.com. Available at: https://www.semrush.com/website/theguardian.com/overview/

Shmargad Y, Coe K, Kenski K, et al. (2022) Social norms and the dynamics of online incivility. *Social Science Computer Review* 40(3): 717–735.

Spadafora N, Farrell AH, Provenzano DA, et al. (2018) Temperamental differences and classroom incivility: exploring the role of individual differences. *Canadian Journal of School Psychology* 33(1): 44–62.

The Guardian (2006) Welcome to comment is free. Available at: https://www.theguardian.com/commentisfree/2006/mar/14/welcometocommentisfree

The Guardian (2009a) Community standards and participation guidelines. Available at: https://www.theguardian.com/community-standards

The Guardian (2009b) Frequently asked questions about community on the Guardian website. Available at: https://www.theguardian.com/community-faqs

The Guardian (2010) Newspaper support in UK general elections. Available at: https://www.theguardian.com/news/datablog/2010/may/04/general-election-newspaper-support

Tice DM, Bratslavsky E and Baumeister RF (2001) Emotional distress regulation takes precedence over impulse control: if you feel bad, do it. *Journal of Personality and Social Psychology* 80(1): 53–67.

Tschantz A, Millidge B, Seth AK, et al. (2023) Hybrid predictive coding: inferring, fast and slow. *PLoS Computational Biology* 19: e1011601.

Vogels EA (2022) Teens and cyberbullying 2022. *Pew Research Center*, 18 December. Available at: https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/

Wood JT (2014) *Interpersonal Communication: Everyday Encounters*. Boston, MA: Wadsworth Cengage Learning.

Zhang Y, Nanduri S, Jiang L, et al. (2023) BiasX: 'Thinking slow' in toxic content moderation with explanations of implied social biases. arxiv.org. Available at: https://aclanthology.org/2023.emnlp-main.300.pdf

Author biographies

Dr Ben Clarke's research focuses on news and political communication, particularly concerning democratic aspects online and environmental communication.

Dr William Hedley Thompson's research focuses on temporal and connected phenomena in behavioural, psychological, and neuroimaging data.