Online Engagement Between Opposing Political Protest Groups via Social Media is Linked to Physical Violence of Offline Encounters

Gallacher, J.D.; Heerdink, M.W.; Hewstone, M.

DOI
10.1177/2056305120984445

Publication date
2021

Document Version
Final published version

Published in
Social Media and Society

SI: Dialogue 2.0: New Perspectives, Enduring Challenges, and Promising Directions'

SM+S
social media + society

Social Media + Society
January-March 2021: 1–16

DOI: 10.1177/2056305120984445
journals.sagepub.com/home/sms
SAGE

Online Engagement Between Opposing Political Protest Groups via Social Media is Linked to Physical Violence of Offline Encounters

John D. Gallacher[1], Marc W. Heerdink[2], and Miles Hewstone[1]

Abstract

The rise of the Internet and social media has allowed individuals with different backgrounds, experiences, and opinions to communicate with one another in an open and largely unstructured way. One important question is whether the nature of online engagements between groups relates to the nature of encounters between these groups in the real world. We analyzed online conversations that occurred between members of protest groups from opposite sides of the political spectrum, obtained from Facebook event pages used to organize upcoming political protests and rallies in the United States and the United Kingdom and the occurrence of violence during these protests and rallies. Using natural language processing and text analysis, we show that increased engagement between groups online is associated with increased violence when these groups met in the real world. The level of engagement between groups taking place online is substantial, and can be characterized as negative, brief, and low in integrative complexity. These findings suggest that opposing groups may use unstructured online environments to engage with one another in hostile ways. This may reflect a worsening of relationships, in turn explaining the observed increases in physical violence offline. These findings raise questions as to whether unstructured online communication is compatible with positive intergroup contact, and highlights the role that the Internet might play in wider issues of extremism and radicalization.

Keywords

group processes, social media, social networks, intergroup contact, political extremism

The rapid expansion of digital communication technologies and the Internet allows individuals to connect and interact in ways that were previously impossible. Today, people can communicate, share ideas, and participate in political discussions from almost anywhere on the globe. This has the potential to have either positive or negative effects on social cohesion and social integration. When in its infancy, it was hoped that the increased interpersonal connection made possible by social media would bring about global expansions in democracy, highlighted by its role in promoting the Arab spring (Howard et al., 2011). Today, this idea has faded, with social media instead seen as posing a fundamental threat to democracy by driving social polarization, disinformation, and hostility (Guess, Barber, et al., 2018; Sunstein, 2017). These opposing optimistic and pessimistic views are difficult to reconcile as the exact relationship between online interactions between groups and offline group dynamics is not known.

In this study, we explore how opposing groups use the online environment to engage spontaneously with one another, what the nature of this engagement is, and whether it relates to group behavior in the real world. Specifically, we investigate whether the degree of inter-group engagement on social media is associated with offline violence between rival groups when they subsequently meet in the real world, and whether the qualities of this conversation moderate such effects. To answer these questions, we analyze conversation data from Facebook event pages preceding 25 recent political protests and rallies in the United Kingdom and the United

[1]University of Oxford, UK
[2]University of Amsterdam, The Netherlands

Corresponding Author:
John Gallacher, Oxford Internet Institute, University of Oxford, Oxford OX1 3JS, UK.
Email: john.gallacher@univ.ox.ac.uk

States, all consisting of a right-wing protest group and a left-wing counter protest. We use a text classifier to estimate the level of outgroup engagement (defined as communication between members of opposing groups), and a combination of established and novel text analysis techniques (natural language processing) to calculate four measures of conversation quality on 73,657 posts within Facebook event pages.

Literature Review

We organize this brief review of the literature around four key topics. First, we consider the relationship between online communication between members of a group and this same group's members' offline group behavior. Second, we consider evidence for the efficacy of online contact in improving relations between groups. Third, we review evidence that communication via the Internet plays a key role in co-radicalization, making both groups involved hold even more extreme worldviews. Finally, we propose how natural language processing can be used to provide a quantitative and objective measure of the types of online conversations between opposing political groups.

The Relationship Between Online Communication and Offline Group Behavior

Online social media activity within groups has been shown to correlate with offline group behavior in cases of social mobilization and political change. Multiple studies have found that increased online activity, often on Facebook and Twitter, has been associated with subsequent increases in protest attendance at a later date. These include the pro-democratic movements of the Arab spring (Steinert-Threlkeld et al., 2015), anti-capitalist and economic inequality protests in the United States and Spain (Bastos et al., 2015), and anti-government protests in Ukraine (Gruzd & Tsyganova, 2015). Digital connectivity has been identified as a driving factor in how these social movements connect, organize, and evolve (Tufekci, 2017), as it reduces the costs of organization and allows activity to erupt in spontaneous and unexpected ways (Enikolopov et al., 2016). In addition, the nature of social media conversations has been shown to be related to the future number of hourly arrests at prolonged one-sided political protests that descend into arson and vandalism (Mooijman et al., 2018); within group conversations which become more moralized may have made this violence appear more socially acceptable (Mooijman et al., 2018), and social media allows both the signaling and gauging of the moral sentiment of others (Barberá et al., 2015). Similarly, violence toward immigrants within western countries has been shown to be related to the degree of anti-refugee sentiment expressed on social media in areas where the violence takes place (Müller & Schwarz, 2018a), while in the United States, anti-Muslim messages disseminated by President Trump over social media correlate with the number of anti-Muslim hate crimes in states where social media usage is high (Müller & Schwarz, 2018b), although the temporal order of online and offline measures here is unclear. These cases show how the online environment may not only affect online behaviors but spreads offline as well.

The Contact Hypothesis Online

Those who espouse the optimistic view that social media can bring about increased social cohesion highlight that by providing new opportunities for individuals from different groups to gather and interact with members from other groups, the Internet could potentially play an influential role in increasing contact, and breaking down barriers between groups (Amichai-Hamburger & McKenna, 2006 ; Schwab & Greitemeyer, 2015). The contact hypothesis proposes that positive face-to-face contact between members of different groups provides one of the best ways to improve relations between these groups if certain facilitating (rather than essential) conditions are met: equal status between groups, the sharing of common goals, intergroup co-operation, personal interaction, and support from authorities (Allport, 1954). The efficacy of offline contact in improving intergroup relations has been demonstrated with groups that differ in terms of, for example, race and religion (K. T. Brown et al., 2003), sexual orientation (Herek & Glunt, 1993), and has been confirmed in experimental (Ioannou et al., 2017) and longitudinal (Ramiah & Al Hewstone, 2013) research and meta-analyses (Davies et al., 2011; Pettigrew & Tropp, 2006). It has also been demonstrated for political views (Sønderskov & Thomsen, 2015); for example, intergroup contact between liberals and conservatives reduced hostility and improves attitudes to opposing parties in the United States (Manbeck et al., 2018), and positive contact with European Union (EU) nationals was associated with support for EU membership during the 2016 Brexit referendum (Meleady et al., 2017). If these same offline effects carry across to the online world, then positive change offline could follow online intergroup contact.

In highly controlled settings, intergroup contact via the Internet has been shown to have positive effects on intergroup relations (White & Abu-Rayya, 2012). This highlights the opportunities offered by the Internet for positive intergroup contact in cases where physical contact is restricted by geographical, political, and economic barriers (Austin, 2006; Hoter, 2009). Amichai-Hamburger and McKenna (2006) go one step further in their "Internet contact hypothesis," and outline not only how all of the conditions for positive intergroup-contact can be satisfied in an online context, but also propose that online contact has an advantage over offline contact as it allows various features to be manipulated to create optimal contact conditions. This hypothesis is echoed in the optimistic view of online interaction promoted by the

world's largest social media platform, Facebook, which announced that it

> is proud to play a part in promoting peace by building technology that helps people better understand each other. By enabling people from diverse backgrounds to easily connect and share their ideas, we can decrease world conflict in the short and long term. (Facebook, 2010)

However, Amichai-Hamburger also warns of the potential risks from misuse of digital platforms, and advocate the careful selection of discussion participants, supervision, and prior agreement to stay on topic and avoid "flaming," which is the act of posting insults, profanity or offensive language with the intention to seek out a negative reaction from the reader (Amichai-Hamburger, 2008). The open and unsupervised nature of social media platforms cannot provide these controls. Social media platforms instead provide the opportunity for unrestrained interactions, often in an anonymous format (users can select their own usernames or profile images, not necessarily linked to their real identity) with very few limitations on the type of language used or moderation of inflammatory content. In such situations, negative rather than positive intergroup contact may occur, and while positive intergroup contact can reduce prejudice, negative intergroup contact may increase it (Graf et al., 2014; Paolini et al., 2010). Negative contact increases the salience of an outgroup individual's group membership, and so any negative effects of contact generalize more strongly to the group as a whole (R. Brown & Hewstone, 2005). While negative intergroup contact is less common than positive intergroup contact in the real world (Graf et al., 2014), this may not be true online. For example, analysis of Facebook groups about the Israeli-Palestinian conflict found little evidence for positive intergroup contact, but rather evidence of hateful antagonistic positions and intolerance (Ruesch, 2011). Within the Facebook pages dedicated to this conflict analyzed by researchers, most content was dedicated toward intragroup mobilization and declaration, and although some pages did self-categorize as "peace groups," with the stated goal of promoting intergroup dialogue, these pages were much less popular than highly partisan pages.

Evidence shows that in large online environments small numbers of communities initiate a large proportion of the intergroup communication, and this communication is often distinctly hostile (Kumar et al., 2018). Social media users in the United States interviewed about interactions with political outgroup members reported that they were stressful and frustrating, and that other users with whom they interacted online were angry and disrespectful (Duggan et al., 2016). This is reflected in evidence that artificial exposure to opposing views on Twitter can increase political polarization (Bail et al., 2018).

Because positive consequences of online engagement with other groups are possible, but certainly not guaranteed, we investigated how naturalistic engagement with outgroup members via social media relates to real-world intergroup behavior. With increases in polarization (Dimock et al., 2014), fragmentation, and extremism (Poushter et al., 2015) throughout the western world, understanding the impact of online connectivity is paramount to inform us about and counter social division. In this study, we provide a unique insight into this question by examining the online relationships between opposing groups at the extremes of the political spectrum. We define this extremism as a belief that ingroup survival is inseparable from a need for hostile action against an outgroup (Berger, 2018a). These hostile actions can vary from discriminatory behavior to verbal attacks or violence (Berger, 2017). Right-wing extremist violence has recently increased across the world (Muhlhausen & McNeill, 2011; Neiwert et al., 2017), and for right-wing extremists the Internet has become a vital tool used to radicalize, recruit, mobilize, and network (Berger, 2018b; Conway & Courtney, 2017). Over time their language has become more aggressive and is associated with a sharp rise in online hate crimes (EUROPOL, 2017). As a result, one of most prominent far-right groups (Mudde, 2007, 2019) in the United Kingdom, Britain First, has recently been banned from the two largest social media sites: Facebook (Facebook Newsroom, 2018) and Twitter (BBC News, 2017). Recognizing the impact of online interactions, Facebook reported that it took this action because the group was sharing hate speech designed to stir up division.

Co-Radicalization and the Internet

It has been suggested that groups from opposite ends of the political spectrum "feed off" one another other in a process of co-radicalization (Knott et al., 2018; Pratt, 2015), a two-way process where different groups reciprocally construct increasingly radicalized worldviews (also referred to as cumulative extremism or mutual radicalization). Often co-radicalizing groups use actions of the other group to justify their own behaviors or prejudices (Ebner, 2017). Offline, this can result in violence by one group being met with violence by the other (Bundesministerium des Innern, 2015).

Evidence for co-radicalization in online spaces is limited, but there is increasing evidence for its occurrence. For instance, areas of the United Kingdom, Germany, Belgium, and France with larger far-right communities and greater anti-Muslim hostility offline also have greater levels of pro-Islamic State content online (Mitts, 2019). This remains the case even when accounting for socio-economic factors such as unemployment and income, suggesting that there may be a link between the offline prejudice and online radicalization.

In addition, there are indications that both far-right and Islamic extremist groups online make sustained references to the other group, with both sides blaming and demonizing the other, and spreading sentiments of victimization and

conspiracy theories about the other group (Fielitz et al., 2018). Similarly, there is evidence that in online spaces overtly hostile language is used to counter hate speech in 39.7% of cases, and this is often met unfavorably—leading to a further reduction in relations rather than the intended improvement (Mathew et al., 2019).

While this evidence suggests that contact between opposing groups online may facilitate co-radicalization, little has been done to research the nature of direct interactions between opposing groups, and it is these between-group interactions which are a key feature of how the co-radicalization process occurs (Moghaddam, 2018). Our study investigates the nature of these direct interactions in an unstructured and relatively unmoderated online space.

For far-right groups, opposition groups such as Muslim communities or anti-fascist counter protestors are often framed as "extreme" and as posing an existential threat to the far-right ingroup to legitimize radical responses (Jackson, 2018). The Internet is moreover becoming recognized as an important facilitating factor in this process (Sirseloudi, 2017), extending the reach of activists, allowing for international cooperation between ideological allies and increasing opportunities for radicalization (Briggs & Strugnell, 2011; Von Behr et al., 2013). This makes far-right groups and counter protest groups an ideal case with which to study the association between online outgroup engagement and digital and real-world group dynamics.

Natural Language Processing

Natural language processing is a fast-growing area of computer science which allows researchers to obtain a quantitative and objective measure of the types of conversations that are happening in online spaces from analysis of the text that is shared (Silge & Robinson, 2017). Measures range from simple sentiment analysis which gives a measure of how positive or negative a comment is, to more advanced constructs which measure the complexity of the language, the specific emotions that are contained, or the level of hostility and aggression.

The Present Research

In the current study, we make use of a number of measurements which are particularly useful when studying intergroup relations and online intergroup conflict. First, we measure sentiment, a broad indication of how positive or negative a post/comment is. Evidence from intergroup contact research shows that positive interaction between opposing groups is more likely to lead to a reduction in group hostility, while negative interaction can have the reverse effect (R. Brown & Hewstone, 2005). Second, we measure integrative complexity (IC), which quantifies the ability of an individual to think and reason with input from multiple perspectives (Streufert & Suedfeld, 1965), and has proven successful in measuring cognitive complexity in situations ranging from international relations and electoral competition to political revolutions (Suedfeld, 2010). The level of IC presented in online communication can provide information about the extent to which authors hold radical or extremist views (Smith et al., 2008), and changes in IC are predictive of international violence (Guttieri et al., 1995; Suedfeld & Bluck, 1988) as well as intergroup conflict (Tetlock et al., 1993). As such, IC may be an important moderator between online outgroup engagement and subsequent improvements in group relations. Finally, we use a measure of online incivility: toxicity. This is defined as a measure of how likely a comment is to make someone leave a conversation, with comments that are defined as being more rude, disrespectful, or unreasonable being more likely to receive a higher "toxicity" score (Wulczyn et al., 2017). This is similar to negative sentiment but goes a step further by including the detection of personal attacks and harassment. As such this is likely to be a useful metric when measuring intergroup communication, as it gives an indication of how antagonistic the communication is.

Here we use these metrics to investigate how opposing groups engage online with one another, and test whether the quantity and quality of their online communications is linked to their behavior when they meet in the real world. Specifically, we look at opposing political groups engaging on Facebook and, based on the idea that hostile communication can further divide already opposed groups, we hypothesize that more communication online will be associated with more violence offline, when members of these groups meet later in the real world. We further predict that the relationship between outgroup engagement and subsequent violence would be moderated by IC, toxicity, and sentiment whereby lower IC, and higher toxicity and sentiment, are associated with greater violence.

Materials and Methods

Ethics

All research was conducted in accordance with the University of Oxford Ethics Committee (Ethics Reference: R55162/RE001). All data collection was conducted using open source methods and publicly available data, and hence, informed consent was not explicitly obtained. No privacy infringements were made, no private groups were joined, and no accounts "befriended" to access data that are not publicly available.

Sampling

Twenty-five physical events were selected for analysis that occurred between October 2015 and October 2017. Each event consisted of a right-wing protest, march or rally that occurred with a corresponding counter-march or protest by the opposing

political side, organized in tandem, on the same day and at the same location. Of these 25 events, 20 occurred in the United Kingdom and five occurred in the United States. Events were selected for the United Kingdom from the most active street-protest groups from each side of the political spectrum, and in the United States, events were selected which occurred in response to the "Unite the Right" rally in Charlottesville in August 2017 (see Supplementary Information, SI, 1).

Once the political events were identified, the conversations taking place online were collected. This was done via the Facebook Graph Application Program Interface (API). We collected conversations taking place on the Facebook pages that were set up to promote the event. This collection method gathered 73,632 comments in total with an average of 1,473 comments per event page, and a total of 2,946 comments on average per event. Once collected, the data were cleaned to remove any conversation that occurred after the planned start time of the event. In doing this we can safely assume that any violence at the event had no impact on conversation online (see SI 1).

Facebook Event Pages are ideal for our study because they are unique in how the online space is linked directly to an offline event, and how they are chronicled and remain accessible for past events. Facebook pages are one of the primary places where online engagement between opposing groups occurs, with the social media platform being used not just for communication within the group but as a primary means for group mobilization. Here, we focus specifically on Facebook event pages, which are the primary method through which groups plan and disseminate information about upcoming marches, protests and rallies, and which, crucially, allow for public discussion and hence for members of differing groups to communicate. While it is possible that other social media platforms may also be used to promote and coordinate these events (such as Twitter), the link between online conversation and offline event on these other platforms is more ambiguous, and inferences about this relation would therefore be more difficult to draw.

Text Analysis Measures

To allow for comparison between events all text analysis measures were coded at the comment level, and then aggregated to the event page level. This resulted in 50 data points for each measure in total (25 right-wing pages, and 25 left-wing pages). These were subsequently aggregated to the event level using a BLUP-based method (Croon & Van Veldhoven, 2007).

Conversation Tone and Sentiment Extremity

Sentiment analysis was performed using RSentiment (Bose, 2017) for *R* (Version 1.1.383, 2017), which classifies each comment into very positive, positive, neutral, negative, very negative categories using a "parts of speech" tagging system. It first classifies each word in the sentence as one of the above categories, and then calculates the overall classification of the comment. To account for negation, the package checks whether each word has been preceded by any negative quantifier and if so, adjusts the score accordingly (Bose et al., 2017).

While this analytic approach does not fully overcome the limitation that sentiment analysis tools lack context as they look at each message in isolation, we also take the average score for the entire conversation for each event page to reduce the impact of individual misclassified messages. From the classification of individual comments, a single "tone" value was calculated for each event page. This tone value ranged on a scale from one to five and was calculated by assigning values from one to five for comments from very negative to very positive and calculating the average score per event page. An event page score of one would represent a page with 100% very negative comments, while a score of five would represent a page with 100% very positive comments. In order to account for the fact that positive and negative sentiments are often not mutually exclusive (Berrios et al., 2015), a sentiment extremity score for each event page was then calculated. This was done by calculating the percentage of comments within the page that were classified as either very positive or very negative. This second measure, ranging from 0 to 100, therefore gives an indication of the emotional extremity of the conversation.

IC

We used an automated IC scoring system, AUTO IC (Gideon et al., 2014; Houck, 2014), to generate IC scores for each comment within an event page, from which the mean IC for the entire conversation on the event page was calculated. The Automated IC system produces a score from one to seven for each comment. This uses the same scoring methodology as human-scored IC. In both systems, scores of one represent a total lack of differentiation (acknowledgment of different viewpoints) or integration (combination/connections of multiple viewpoints). Scores from two to three represent levels of increasing differentiation, but no integration. Scores from four to six represent increasing and moderate to high levels of differentiation and integration. A score of seven indicates high differentiation plus high integration.

Individuals who display higher IC tend to construct more accurate and balanced perceptions of other people, use more information when making decisions, as well as holding less extreme views, and as a result these individuals are shown to be less prejudiced and are better able to resolve conflicts cooperatively with outgroup members. Furthermore, within-group discussions with higher levels of IC have been shown to decrease displays of greed and fear, and reduce the likelihood that a group would decide to take a competitive stance against others (Park & DeShon, 2018). Recently, AUTO IC has been used successfully for the study of online terrorist

content, demonstrating the validity of the application to the digital domain (Houck et al., 2017).

Toxicity

We used the Google Perspective API (Google Project Jigsaw, 2018) to measure the level of toxicity within the online conversations. This classification tool was designed by Google's "Project Jigsaw" and "Counter Abuse Technology" teams with the aim of promoting better discussions online (Wulczyn et al., 2017). The model gives a toxicity score for each comment on a scale ranging from zero (least toxic) to one (most toxic). In the current study, each comment was sent through the Perspective API, and from this the average toxicity rating for each event page calculated.

Outgroup Engagement

In order to identify occurrences of outgroup engagement we trained a neural network (Sebastiani, 2000) to classify comments as either "within-group," for comments that were directed toward other ingroup members, or "between-group," for comments that were directed toward a member of the outgroup. A "between-group" comment could therefore be either a member injecting a comment into the event page of the opposing group, or a reply to this injection from a member of the incumbent page. The proportion of between-group comments for each event page was calculated (*s*ee SI 5 for further details of how this measure was operationalized). We define this type of communication between members of opposing groups as "outgroup engagement" as it falls short of the level of interpersonal involvement and connection required for traditional intergroup contact but shares some important characteristics with it.

The neural network was trained on a set of 1,000 randomly sampled comments from the overall dataset to ensure that group-specific language and idioms were accurately interpreted, as such elements may be misinterpreted by generic lexicon measures (Omand et al., 2012). Each comment in the training set was human coded. The default coding option was within-group communication, and this was selected in all cases where a decision could not be made (either through a lack of information or clarity). To ensure accuracy of the human coding, all comments were coded by a second coder who was blinded to the hypotheses, and intercoder reliability (ICR) scores calculated. For the training set the ICR was 97.80% with a Scott's PI of .96. For the test set the ICR was 95.90% with a Scott's PI of .90 (see SI 5).

The overall accuracy of the classifier was 89.0%, with a sensitivity of 85.9% and a specificity of 89.9% when checked against representative test set of 1,000 comments. This is therefore a conservative judgment classifier with regard to outgroup engagement classification, reflecting the conservative nature of setting within-group conversation as the default.

As an additional measure of classification consistency, we calculated the proportion of comments for each user within an event page that are given the same label by the classifier (as either within-group or between-group communication). Overall, we found 91.0% consistency in these ratings. We judged this to be a high level, especially as we would not expect consistency to reach 100%; "home" users with the event page started by their own ingroup should be classified differently when replying to an outgroup member that visits a page compared to when replying to an ingroup member on the same event page.

Similar approaches using machine learning to classify digital text have previously been shown to be valid with regard to online comment abuse detection (Chu et al., 2017), machine translation (Wołk & Marasek, 2015), and sentiment analysis (Kim, 2014), but to the best of our knowledge, this is first time such methods have been used to identify cases of online intergroup engagement.

Violence

We developed two measures of violence for each real-world event. In all cases, this was based on open source intelligence taken from a range of sources, including professional journalistic reports, citizen journalism, photos, videos, and police reports on arrest statistics (See SI 2). The first violence measure is a binary "absence or presence of violence" (0 = *absent*, 1 = *present*) based on whether reports stated that violence occurred at the event. The second measure allocates a degree of violence score based on seven security industry-standard violence indicators. We fit a latent trait model to these indicators, which assumes that each event has an unobserved (latent) true level of violence that manifests itself in the absence or presence of these indicators. The model determines two parameters for each indicator: the "severity," reflecting the level of violence at which this indicator is likely (>50%) to be present, and "discrimination," which reflects the sensitivity of the indicator to changes in violence. These parameters are then used to estimate the latent violence score for each event on a continuous scale (see SI 2).

We checked the robustness of this measure using sensitivity analysis, whereby each indicator was removed in turn and the analysis repeated. The results were very similar in all cases, suggesting that no individual indicator is responsible for the observed effects (see SI 4).

Statistical Methods

All statistical analyses were conducted in R. For each model, the optimum combination of predictors (text analysis measures) was selected using Akaike Information Criterion (AIC). We tested the predictive power of text analysis measures, including outgroup engagement, on violence using a logistic regression (generalized linear model, GLM) (SI 2). This analysis was then repeated replacing presence/absence

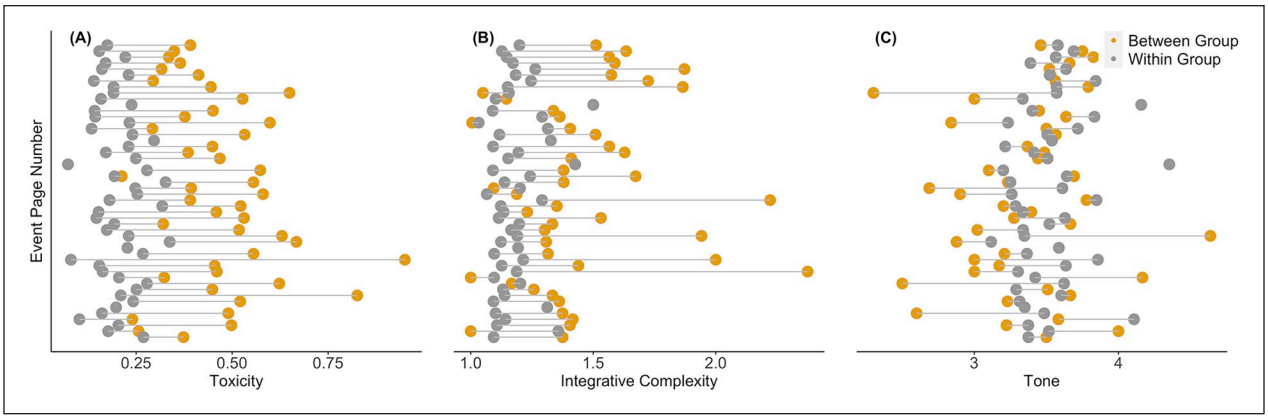

Figure 1. The differences in conversation qualities for between-group conversation and within-group conversation for each event page. Dumbbell plots show that (a) toxicity is higher in between-group conversation, (b) integrative complexity is higher in between-group conversation, and (c) there is no difference in Tone for between-group communication and within-group conversation.

of violence with degree of violence, this time using a linear model (LM). To test for moderation effects of the quality of both within-group and between-group conversations on violence measures we tested for interaction effects using a GLM and LM. We used paired *t*-tests to compare the nature of the comments in between-group and within-group communication within the same event.

The total conversation size within the event page discussion (number of comments) was not found to explain any variance and therefore was not included in these models. All events contained a degree of intergroup contact; however, five event pages (representing approximately half the event conversation in each case) contained no intergroup contact. Therefore, when comparing the qualities of the intergroup contact with subsequent violence for these events we used the values of the event page that did contain contact and did not aggregate across pages. When comparing the length of continuous chains of comments, a non-parametric Mann–Whitney U independent samples test was performed to account for a negatively skewed nature of the comment chain length distributions.

The models were shown to be robust through the absence of influential data points and multicollinearity (SI 4). Where we detected multicollinearity (testing for moderation effects) we resolved this by centering the predictors prior to analysis. Throughout the results all measures are shown as estimate ± *SE*.

Results

Overall, 32.0% of comments across all event pages were classified as outgroup engagement. To understand better what type of outgroup engagement occurred, we compared the between-group conversation on an event page to the within-group conversation on the same page. We found that compared to within-group conversation, between-group conversation displayed a higher level of toxicity (paired *t*-test, between-group $M = 0.47 \pm 0.02$, within-group $M = 0.20 \pm 0.01$, $df = 44$, $t = 12.36$, $p < .001$, $d = 1.84$) and a higher level of IC (paired *t*-test, between-group $M = 1.46 \pm 0.04$, within-group $M = 1.18 \pm 0.01$, $df = 44$, $t = 6.95$, $p < .001$, $d = 1.04$). With regard to tone, between-group conversation was slightly more negative than within-group conversation, but this difference was not significant (paired *t*-test, between-group $M = 3.37 \pm 0.07$, within-group $M = 3.52 \pm 0.04$, $df = 44$, $t = -1.89$, $p = .065$, $d = 0.28$) (Figure 1). In addition, to identify whether there was a difference in the duration of between-group conversations and within-group conversations, we calculated the average length of a continuous chain of between-group comments, and found this was shorter than the average length of a continuous chain of within-group comments (Wilcoxon signed rank test, between-group: 2.58 ± 0.02, within-group: 3.44 ± 0.03, $U = 45{,}144{,}000$, $p < .001$). This suggests that outgroup engagement often consists of short-lived interjections in the other group's discussions, which invite a prompt response.

To test whether conversation qualities and the proportion of outgroup engagement with a Facebook event page are associated with offline violence, we first measured these variables separately on the left-wing and right-wing pages relating to each event (*see SI 7, Figure S4 & Table S7, for differences in conversation metrics between right-wing and left-wing pages*), and then aggregated these scores to the event level by calculating best linear unbiased predictors (BLUPs); we then tested their association with offline violence using a logistic regression (*SI 4*). We used a stepwise method to compare the ability of different combinations of variables assessing the nature of online conversations to statistically predict offline violence later in time, and selected the best model based on AIC.

Real-word violence was associated with two conversational variables: the level of outgroup engagement (i.e., the proportion of between-group conversation on a page) and the IC of the conversations. Violence was more likely if conversations previously had higher levels of outgroup engagement

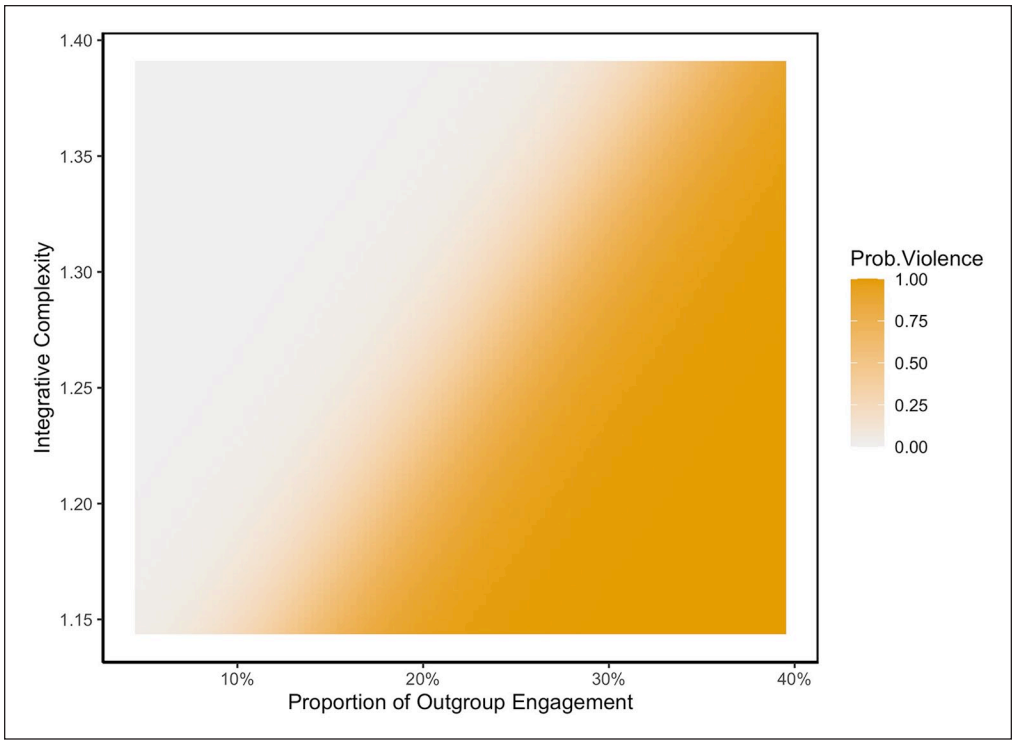

Figure 2. Probability of an event becoming violent for levels of outgroup engagement and Integrative Complexity in preceding conversations on Facebook event pages. Offline violence is more probable when there is more online outgroup engagement prior to the event, or when the integrative complexity of the online discussion is lower.

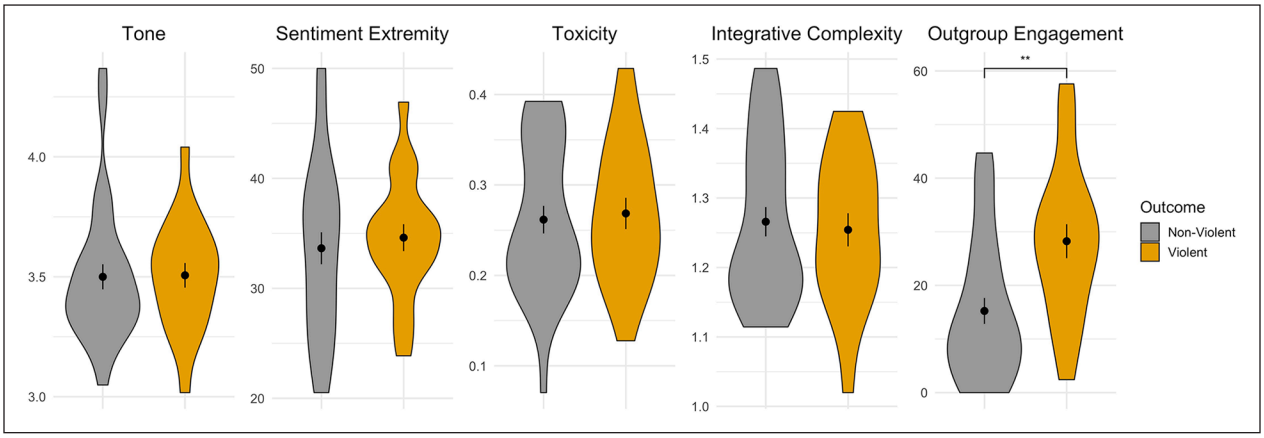

Figure 3. Comparison of violent versus non-violent conversation qualities. Differences in Tone, Sentiment Extremity, Toxicity, Integrative Complexity and outgroup engagement occurring on Facebook event pages for events which subsequently became violent versus those which remained peaceful. Mean and standard error are given by dots and lines, respectively. ** denotes $p < .001$.

and lower levels of IC (Figure 2) (GLM, $n = 25$, outgroup engagement: $B = 0.38 \pm 0.16$, Wald's $z = 2.33$, $p = .020$; IC, multiplied by 10 to account for the limited range [1.13–1.39 on a 1–7 scale]: $B = -3.20 \pm 1.60$, Wald's $z = -2.00$, $p = .046$; notation: estimate $\pm$ *SE*). In other words, each 1.0% increase in the proportion of outgroup engagement on an event page increased the odds of the event becoming violent by a factor of 1.46, while each 0.1 unit increase in the average conversation IC decreased the odds of violence occurring by a factor of 24.50. Model selection did not retain tone, sentiment extremity or toxicity as variables associated with subsequent violence. Figure 3 illustrates this result, such that events which became violent displayed higher levels of outgroup engagement (Welch two sample *t*-test, violent $M = 28.25 \pm 2.71$, non-violent $M = 15.23 \pm 2.12$, $df = 24$, $t = -3.78$, $p = .001$, $d = 1.56$) and lower levels of IC than events which remained

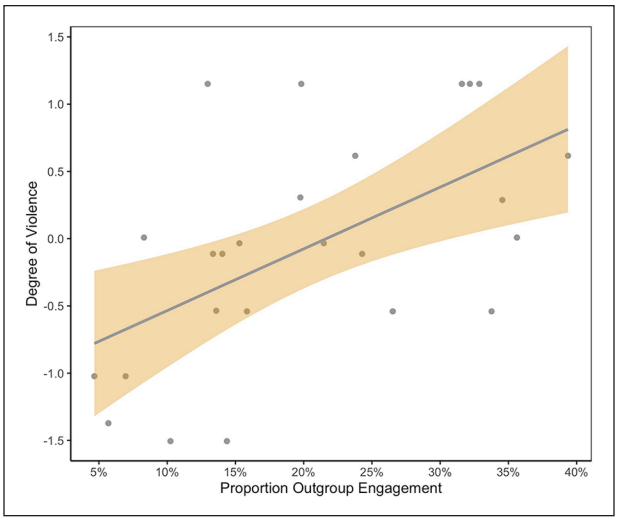

Figure 4. Outgroup engagement and degree of subsequent violence. The level of outgroup engagement in Facebook event pages preceding a political protest is associated with the degree of violence at that event.

peaceful (Welch two sample *t*-test, violent $M=1.27\pm0.02$, non-violent $M=1.25\pm0.02$, $df=24$, $t=0.40$, $p=.70$, $d=0.17$), although the latter is only significant in the full GLM.

To determine whether the two variables, outgroup engagement and IC, were also associated with the *degree* of violence, we developed a continuous measure of violence based on the standard indicators of violence used in the security industry (see SI 2) and repeated the analysis. In the best model according to AIC, only outgroup engagement was significantly statistically associated with the degree of violence, with more outgroup engagement being associated with more violence (Figure 4, LM, $n=25$, $B=0.05\pm0.01$, $t_{23}=3.30$, $p=.003$). The distribution of these violence metrics across event pages is shown in the Supplementary Information (SI) Table S3 and Table S4.

We hypothesized that the relationship between outgroup engagement and subsequent violence would be moderated by IC, toxicity, and tone. We therefore tested whether any interactions between outgroup engagement and these metrics in online conversations were associated with the degree of offline violence. We found no significant interactions between outgroup engagement and IC, and outgroup engagement and tone (LM, $n=25$, outgroup engagement×IC; $B=-0.29\pm0.24$, $t_{21}=-1.24$, $p=.230$, outgroup engagement×Tone $B=0.07\pm0.13$, $t_{21}=0.52$, $p=.607$), but a significant negative interaction between outgroup engagement and toxicity (outgroup engagement×toxicity; $B=-0.81\pm0.38$, $t_{21}=-2.15$, $p=.044$). However, this latter model did not account for more variance in the degree of offline violence than outgroup engagement in isolation. We also tested whether interactions between outgroup engagement and conversation quality were associated with the presence rather than degree of violence, and found that they were not (GLM, $n=25$, outgroup engagement×IC; $B=-0.28\pm1.20$, Wald's $z=0.24$, $p=.814$, outgroup engagement×Toxicity; $B=-2.74\pm2.21$, Wald's $z=-1.24$, $p=.215$, outgroup engagement×Tone; $B=-0.18\pm0.68$, Wald's $z=0.26$, $p=.792$). To confirm that the quality of within-group conversations was not masking the role of quality of between-group conversations (outgroup engagement), we tested for interactions between the quality of only the between-group conversations for each event and degree of violence, and again found no evidence for moderation (LM, $n=25$, outgroup engagement×between-group IC; $B=0.002\pm0.08$, $t_{21}=-0.03$, $p=.980$, outgroup engagement×between-group toxicity; $B=-0.22\pm0.20$, $t_{21}=-1.15$, $p=.265$, outgroup engagement×between-group tone; $B=-0.005\pm0.06$, $t_{21}=-0.10$, $p=.923$). Thus, our hypothesis that conversation quality moderates the effect of outgroup engagement was not supported.

Given the negative relationship between IC and the occurrence of violence, and the higher level of IC in between-group conversation compared to within-group conversation, we investigated if either the IC of within-group conversations or the IC of the between-group conversations was more strongly associated with subsequent violence. We tested this while accounting for the level of outgroup engagement, and found that only the IC of the within-group conversations was associated with the presence/absence of violence (within-group IC $B=-7.54\pm3.60$, Wald's $z=-2.10$, $p=.036$, between-group IC $B=0.07\pm0.43$, Wald's $z=0.17$, $p=.864$). In fact, replacing the IC of the overall conversation with IC of the within-group conversations improved the association between the model's specified variables and violence ($\Delta AIC=-7.70$).

Discussion

In this study, we examined whether the frequency and quality of naturally occurring online conversations between opposing political groups on Facebook is associated with offline physical violence at subsequent real-world events. We found that the level of outgroup engagement on Facebook event pages was the variable most consistently associated with both the presence and degree of violence during a subsequent encounter between the groups. Overall, we found that the Facebook event pages used by groups to mobilize and gather support for a march or rally are a place where online outgroup engagement occurs, with 32.0% of the overall conversation across all pages occurring between members of opposing groups. We consider this a substantial percentage, given that previous studies estimated that within right-leaning communities on Twitter up to 93.0% of interactions occur between ingroup members (Conover et al., 2011). The extent of this communication between groups, and the societal impact of the violence associated with this communication, emphasizes the importance of studying such online conversations.

Conversation Quality and Subsequent Offline Violence

In addition to the quantity of communication between groups, our findings show that conversation quality is also associated with future violence. Specifically, lower levels of IC in the conversation were associated with violence during the offline encounter. This aligns with previous findings that linked decreases in IC with the deterioration of group relations, ranging from more competitive intergroup behavior (Park & DeShon, 2018) to international violence (Suedfeld & Bluck, 1988). Individuals who display higher IC tend to construct more accurate and balanced perceptions of others, use more information when making decisions, and hold less extreme views. As a result, these individuals are shown to be less prejudiced and are better able to resolve conflicts cooperatively with outgroup members (Tetlock et al., 1993). While high IC is not traditionally held as one of the conditions required for, or mediators of, positive intergroup contact, it measures the ability to think and reason with input from multiple perspectives, and this ability to take others' perspective and empathize with them is a key mediator of how contact improves outgroups attitudes (Pettigrew & Tropp, 2008). It is possible that in an online environment stripped of much individuating and subtle information the ability to overtly demonstrate multiple viewpoints becomes critical.

When comparing the between-group and within-group communication on an event page, we found that between-group communication was more toxic, indicating that it is more rude, aggressive, or disrespectful to the outgroup. Interestingly, this same communication was also higher in IC, suggesting that while this communication is quite negative, it also engages with opposing views to a greater extent than when ingroup members speak with each other. Furthermore, we found that the IC of the between-group communication was not associated with greater violence, but rather it was the IC of the within-group communication which was, and indeed to a greater extent than the IC of the overall conversation. This suggests that the IC of the conversations taking place between ingroup members is most directly related to group behavior. It may be that less complex conversations reflect an increased homogeneity of the ingroup, and an increased clarity concerning norms regarding interaction with the outgroup, enhancing group identification, and perhaps increasing the likelihood that a group may be provoked or respond to inflammatory triggers in a group fashion rather than in an individual manner. It should also be noted though that as these conversations are taking place in a public forum, it is possible for outgroup members to "observe" the opposing groups' ingroup conversations and this may affect behavior. Together, these findings suggest a dynamic in which different aspects of between-group and within-group communication reflect a group's disposition to engage in outgroup-directed violence.

Nature of Outgroup Engagement in the Conversation

The benefits of intergroup contact are premised on such contact being positive (R. Brown & Hewstone, 2005); in the current study, however, the outgroup engagement could not be characterized as such. We found that naturally occurring communication between members of opposing groups was more toxic than equivalent within-group conversations, suggesting that positive experiences would not be felt by either group involved in these exchanges. This corroborates previous findings that online discussions between ideologically opposed communities typically carry a negative sentiment (H. T. P. Williams et al., 2015). In addition, across all event pages, the average length of a continuous chain of conversation between members of opposing groups was significantly lower than for within-group conversations, indicating shorter instances of intergroup than intragroup contact. While some interactions were longer, on the whole the interactions are far too short and fleeting for any level of personal or prolonged contact to occur. The short nature of the conversations suggests that those taking part are not motivated to maintain the conversation for long. This might reflect the negative nature of the conversation pushing participants away, or it could reflect a fact of the social network platform itself promoting short-term conversations in a constantly changing and updating digital environment. Short-term exposure does not prevent negative, stereotype reinforcing contact from occurring (MacInnis & Page-Gould, 2015). In addition, it requires more time to develop positive impressions online than offline (Jarvenpaa & Leidner, 1999; Walther, 1996) and so these short exchanges, even when positive, may be too fleeting to lead to positive group outcomes such as a reduction in prejudice and discrimination, and an improvement in relations. It should, however, be noted that these findings do not take into account that the same individual may take part in multiple exchanges. Because we anonymized the dataset and looked solely at the content of messages sent, more developed exchanges may be occurring over time in a number of short bursts of engagement.

The relative anonymity of users, due to a lack of individuating information beyond username and profile image being provided, may have been a further obstacle to positive effects of intergroup engagement (Islam & Hewstone, 1993; Lee, 2007); indeed, anonymity has previously been found to reduce the positive effects of computer-mediated intergroup contact under controlled conditions (Schumann et al., 2017). The abrasive and confrontational nature of the discussions (demonstrated by the high toxicity) may instead have increased the salience of group memberships (an element which could be tested in future research). This salience may lead those communicating online to generalize their predominantly negative experience more strongly to the outgroup as a whole, increasing intergroup anxiety, promoting negative stereotypes and damaging chances for future positive interactions (R. Brown

& Hewstone, 2005). In the absence of face-to-face cues and prior personal knowledge about other members of the conversation, then whatever subtle social cues do appear in the online environment take on a much larger weight (Bacev-Giles & Haji, 2017; Postmes et al., 1998). This combination of highly toxic interactions that are short-lived and with low individuality, but highly salient group membership, is a likely explanation for why online outgroup engagement in confrontational situations is associated with negative group outcomes, in this case an increase in offline violence.

Criticism of social media dividing societies has often cited the potential for these platforms to create ideological "echo chambers" (Bright, 2018; Conover et al., 2011)—networks of like-minded people who confirm each other's opinions instead of promoting critical thought. Furthermore, these criticisms assume that increasing digital connection and "breaking down echo chambers" such that individuals interact with people from other social groups will naturally lead to positive outcomes (Berke, n.d.). Our findings, however, are not only at odds with the notion of echo chambers—with 32.0% of all communication taking place between groups—but also challenge the assumption that breaking down echo chambers will necessarily improve intergroup relations. Instead, our results align with findings from the offline domain in showing that such improvements may not occur unless at least some of the key conditions for positive intergroup contact are met (Allport, 1954). A growth of recent evidence also suggests, like our findings, that online echo chambers may not be occurring as commonly as expected (Dubois & Blank, 2018; Guess, Lyons, et al., 2018; O'Hara & Stevens, 2015); however, evidence is limited that online intergroup exposure, such as it is, is associated with improvements in group relations (Yardi & Boyd, 2010). Our results provide evidence that may help to explain this apparent paradox. In our sample, online engagement with the outgroup is occurring, but its limited quantity and predominantly negative quality is unlikely to promote positive group outcomes and reduce antagonism between the groups.

Adversarial Nature of Opposing Groups

Given our focus on communication between protest groups and counter-protest groups, the starting point of the contact situation was likely to be adversarial by default, and outcomes of online intergroup contact may be different with more benign or neutral initial positions (Gehlback et al., 2018). Moreover, the online contact environment is prone to attract individuals with stronger outgroup prejudices (Hasler & Amichai-Hamburger, 2014), and given that the structure of online networks facilitates ingroup contact, engaging the opposing side in discussion may well require, or at least be typically associated with, a motivation to engage the outgroup in online intergroup conflict. This competitiveness, however, is a characteristic feature of intergroup relations (Wildschut et al., 2003), and, combined with the fact that online contact is primarily text based (as in the present study, via Facebook pages), which can itself increase the chance of conflict (Schroeder et al., 2017), conflictual online communication—between left-wing and right-wing political groups, or otherwise—is unlikely to be rare. The societal importance of the outcomes studied here (including, in some cases, bodily harm) highlights how important it is to study the association between online intergroup contact and its behavioral correlates in the real world.

It should be noted that some of the observed adversarial conversation may come from "troll" accounts who act deliberately to inflame or provoke other members of the conversation. Indeed, this type of behavior from inauthentic accounts run by state proxies has been shown to lead to a worsening of online conversations, including an increase in the level of toxicity and a reduction in the IC (Gallacher & Heerdink, 2019). In the current study, we cannot differentiate between genuine social media users and inauthentic accounts, and it is therefore possible that some toxic outgroup engagement was a result of this type of behavior. However, this does not affect our results, because regardless of whether an outgroup provocation is issued from a real or troll account, the effect on the recipient in their perception of the outgroup remains the same.

Limitations and Future Directions

The main limitation of this study is that the evidence we have reported does not allow us to demonstrate that this relationship between online and offline behavior is causal. There may be wider events which are driving both the increase in online hostility and subsequent offline violence in parallel. These variables may include the wider media environment and a specific focus on far-right related issues, political activity, and key leader expressions, as well as highly relevant real-world events such as terror attacks—which have been shown previously to lead to spikes in far-right online activity and hate speech (M. L. Williams et al., 2019). Regardless of this limitation, we believe that our findings are novel and useful and may be indicative of a wider trend where antagonistic online discussions are associated with offline actions later in time. In this sense, the online activity may be viewed as a measure of the "temperature" or "atmosphere" of the group relations at any given time, which only expresses itself physically when the groups subsequently meet in the real world. Future experimental research, which would be ethically demanding in this sphere, could demonstrate that the relationship we found between the online and offline worlds is a causal one. In the absence of such experimental data, data such as we have analyzed (where online conversations link with, and temporally precede, the offline behavior at the relevant event) can make useful contributions toward a better understanding of these issues.

One important and related question to answer in this regard is whether the same people were taking part in the

online and offline conversations. For ethical reasons, we did not store the names or profiles of those taking part in the online conversations, and we can therefore provide no insights on this. However, given that we studied event pages aiming to coordinate the offline rallies, it is likely that a significant proportion of the online users were also present offline. Besides, for our general argument, it is just as important whether more extreme online exchanges are associated with more violence offline by the same *or other* participants. While having more direct evidence in this regard would help further our understanding of how online interaction can translate into offline activities at the individual level, this is not what our study aimed to achieve, and it is difficult to foresee how such research could be done while respecting individual privacy. Instead, we aimed to study group-level effects, acknowledging that the individuals within groups varied in the extremity of their views, and that different individuals will partake at different times, and to test whether groups which had certain online conversation characteristics as a whole were more prone to be involved in more violent events in the real world.

In addition, as is the case with field studies of intergroup contact, there is a self-selection bias with the participants taking part in the conversations. As discussed above, those who are willing to partake in contact with the opposing group may hold stronger outgroup prejudices (Hasler & Amichai-Hamburger, 2014). Forced intergroup contact has been shown to have larger effects than voluntary contact (Pettigrew & Tropp, 2006); in the current study, it can be thought that the individual "reaching out" to the opposing member is making a voluntary engagement, while the recipient is having this interaction forced upon them. Whether effects of either voluntary or forced contact exist within online environments remains to be seen. Equally, our results are restricted to political extremism in the United Kingdom and United States; future research should seek to replicate our main findings in other countries, as well as in domains other than right- and left-wing politics (e.g., hooliganism, separatist conflicts, or racial divides).

Future research could also focus on identifying cases where positive online intergroup contact does occur, and how to generalize these conditions to the wider ecosystem, as well what structural changes could be made to the current online environments in order to encourage more developed, sustained and positive contact between members of opposing groups. These structural changes could include reducing the level of anonymity such that positive interactions with outgroup members in non-political conversations (which are more diverse (Barberá et al., 2015) carry across to political conversations. Alternatively, algorithmic additions, which a user can control, which suppress hostility and promote civility, may help to counter evidence that moral outrage and aggression spread faster on social media than positive content (Brady et al., 2017; Crockett, 2017) and help to rebalance the perceived hostility of outgroup members (Duggan et al., 2016).

Conclusion

We provide evidence that for adversarial political protest groups online conversations are associated with subsequent offline group behavior. We show that for highly charged issues spontaneous engagement with members of opposing groups is fairly frequent, but that social media platforms are failing to facilitate positive outgroup engagement between these antagonistic groups. In fact, the style and nature of such online exchanges is more indicative of negative rather than positive intergroup contact. The superficial and hasty nature of group interactions is likely to reinforce pre-existing prejudices, generate negative affective states and even lead to a situation where groups co-radicalize and polarize their views through unsavory contact with one other. Our results suggest that Facebook was misguided in the claim that it is decreasing conflict simply through enabling the connection of individuals from diverse backgrounds, at least in the case of political groups.

Exposure to uncivil disagreements online is associated with negative effects, including withdrawal and isolation from online conversations (Bode, 2016), increased perception of social distance between groups (Iyengar et al., 2012), and increased affective polarization (Suhay et al., 2018). We take this work a step further and demonstrate that uncivil disagreements between groups on social media are associated with, and can statistically predict, violence when these groups meet in the real world. This is not to say that online intergroup contact and digital communication as a whole cannot lead to positive intergroup outcomes, but rather that the "natural environment" that currently exists within social networking sites is not conducive to nurturing it, especially for those at the extremes of a social, and in this case political, spectrum.

Acknowledgements

We thank members of the Oxford Center for the Study of Intergroup Conflict for their helpful feedback.

Author Contributions

J.D.G. conceived the study, collected the data, and developed the inter-group comment classifier. J.D.G. and M.W.H. developed the statistical models and analyzed the data. J.D.G, M.W.H, and M.H wrote the manuscript. We thank members of the Oxford Centre for the Study of Intergroup Conflict for their helpful feedback.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by grants from EPSRC and the University College Oxford Radcliffe Scholarship. The second author's contribution to this project was supported by a grant from the Netherlands

Organisation for Scientific Research (NWO 446-16-015). The funding bodies played no further role in designing or implementing the research, and the authors declare no competing interests.

ORCID iD

John D. Gallacher https://orcid.org/0000-0003-1292-8079

Supplemental Material

Supplemental material for this article is available online.

References

Allport, G. (1954). *The nature of prejudice*. Addison-Wesley Publishing Company. https://doi.org/10.1002/9780470773963

Amichai-Hamburger, Y. (2008). The contact hypothesis reconsidered: Interacting via internet: Theoretical and practical aspects. *Psychological Aspects of Cyberspace: Theory, Research, Applications*, 209–227. https://doi.org/10.1017/CBO9780511813740.010

Amichai-Hamburger, Y., & McKenna, K. Y. A. (2006). The contact hypothesis reconsidered: Interacting via the internet. *Journal of Computer-Mediated Communication*, *11*(3), 825–843. https://doi.org/10.1111/j.1083-6101.2006.00037.x

Austin, R. (2006). The role of ICT in bridge-building and social inclusion: Theory, policy and practice issues. *European Journal of Teacher Education*, *29*(2), 145–161. https://doi.org/10.1080/02619760600617284

Bacev-Giles, C., & Haji, R. (2017). Online first impressions: Person perception in social media profiles. *Computers in Human Behavior*, *75*, 50–57. https://doi.org/10.1016/j.chb.2017.04.056

Bail, C., Argyle, L., Brown, T., Bumpus, J., Chen, H., Hunzaker, M. B., . . .Volfovsky, A. (2018). Exposure to opposing views can increase political polarization: Evidence from a large-scale field experiment on social media. *Proceedings of the National Academy of Sciences of the United States of America*, *118*, 9216–9221. https://doi.org/10.17605/OSF.IO/4YGUX

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting From left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Bastos, M. T., Mercea, D., & Charpentier, A. (2015). Tents, tweets, and events: The interplay between ongoing protests and social media. *Journal of Communication*, *65*(2), 320–350. https://doi.org/10.1111/jcom.12145

BBC News. (2017, December 18). *Twitter suspends Britain First leaders*. http://www.bbc.co.uk/news/technology-42402570

Berger, J. M. (2017). *Extremist construction of identity: How escalating demands for legitimacy shape and define in-group and out-group dynamics. Terrorism and Counter-Terrorism Studies*. The International Centre for Counter—Terrorism—The Hague. https://doi.org/10.19165/2017.1.07

Berger, J. M. (2018a). *Extremism*. The MIT Press.

Berger, J. M. (2018b). *The alt-right Twitter census: Defining and describing the audience for alt-right content on Twitter*. VOX-Pol Network of Excellence. https://www.voxpol.eu/new-research-report-the-alt-right-twitter-census-by-j-m-berger/

Berke, J. (n.d.). *Mark Zuckerberg says the world is much more divided than he ever expected*. World Economic Forum. https://www.weforum.org/agenda/2018/02/mark-zuckerberg-says-he-thought-facebook-could-solve-a-lot-of-problems-but-the-world-is-more-divided-than-he-expected

Berrios, R., Totterdell, P., Kellett, S., & Brose, A. (2015). Eliciting mixed emotions : A meta-analysis comparing models, types, and measures. *Frontiers in Psychology*, *6*, 1–15. https://doi.org/10.3389/fpsyg.2015.00428

Bode, L. (2016). Pruning the news feed: Unfriending and unfollowing political content on social media. *Research & Politics*, *3*(3), 1–8. https://doi.org/10.1177/2053168016661873

Bose, S. (2017). *Package RSentiment.'* https://cran.r-project.org/web/packages/RSentiment/RSentiment.pdf

Bose, S., Saha, U., Kar, D., Goswami, S., Nayak, A. K., & Chakrabarti, S. (2017). Rsentiment: A tool to extract meaningful insights from textual reviews. In S. Satapathy, V. Bhateja, S. Udgata, & P. Pattnaik (Eds.), *Advances in intelligent systems and computing* (*Vol. 516*, pp. 259–268). Springer. https://doi.org/10.1007/978-981-10-3156-4_26

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Briggs, R., & Strugnell, A. (2011). *Radicalisation: The Role of the Internet. A working paper of the PNN*. Institute for Strategic Dialogue. https://www.isdglobal.org

Bright, J. (2018). Explaining the emergence of echo chambers on social media: The role of ideology and extremism. *Journal of Computer-Mediated Communication*, *23*, 17–33. https://doi.org/10.2139/ssrn.2839728

Brown, K. T., Brown, T. N., Jackson, J. S., Sellers, R. M., & Warde, M. J. (2003). Teammates on and off the field? Contact with black teammates and the racial attitudes of white student athletes. *Journal of Applied Social Psychology*, *33*, 1379–1403. https://doi.org/10.1111/j.1559-1816.2003.tb01954.x

Brown, R., & Hewstone, M. (2005). An integrative theory of intergroup contact. *Advances in Experimental Social Psychology*, *37*, 255–343. https://doi.org/10.1016/S0065-2601(05)37005-5

Bundesministerium des Innern. (2015). *2015 Annual report on the protection of the constitution. Facts and trends*. https://www.verfassungsschutz.de/embed/annual-report-2015-summary.pdf

Chu, T., Jue, K., & Wang, M. (2017). *Comment abuse classification with deep learning*. Stanford University. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2762092.pdf

Conover, M., Ratkiewicz, J., & Francisco, M. (2011). Political polarization on twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (pp. 89–96). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2847/3275

Conway, M., & Courtney, M. (2017). *Violent extremism and terrorism online in 2017: The year in review*. VOX-Pol Network of Excellence. https://www.voxpol.eu/vox-pol-year-review-published/

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*, 769–771. https://doi.org/10.1038/s41562-017-0213-3

Croon, M. A., & Van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, *12*(1), 45–57. https://doi.org/10.1037/1082-989X.12.1.45

Davies, K., Tropp, L. R., Aron, A., Pettigrew, T. F., & Wright, S. C. (2011). Cross-group friendships and intergroup attitudes: A meta-analytic review. *Personality and Social Psychology Review*, *15*(4), 332–351. https://doi.org/10.1177/1088868311411103

Dimock, M., Kiley, J., Keeter, S., & Doherty, C. (2014). *Political polarization in the American public*. https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/

Dubois, E., & Blank, G. (2018). The echo chamber is overstated: The moderating effect of political interest and diverse media. *Information Communication and Society*, *4462*, 1–17. https://doi.org/10.1080/1369118X.2018.1428656

Duggan, M., Smith, A., & Page, D. (2016). *The political environment on social media*. Pew Research Centre. http://www.pewinternet.org/2016/10/25/the-political-environment-on-social-media/

Ebner, J. (2017). *The rage: The vicious circle of Islamist and far-right extremism*. I.B.Tauris.

Enikolopov, R., Makarin, A., & Petrova, M. (2016). *Social media and protest participation: Evidence from Russia to understand whether social media indeed promotes protest participation*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2696236

EUROPOL. (2017). *European Union terrorism situation and trend report (EU TESAT) 2017*. https://doi.org/10.2813/237471

Facebook. (2010). *Peace on Facebook*. https://www.facebook.com/helppeople/posts

Facebook Newsroom. (2018). *Taking action against Britain first*. https://newsroom.fb.com/news/h/taking-action-against-britain-first/

Fielitz, M., Ebner, J., Guhl, J., & Quent, M. (2018). *Loving hate. Anti-Muslim extremism, radical Islamism and the spiral of polarization*. https://www.isdglobal.org/isd-publications/hassliebe-muslimfeindlichkeit-islamismus-und-die-spirale-gesellschaftlicher-polarisierung-deutsch/

Gallacher, J. D., & Heerdink, M. W. (2019). Measuring the effect of Russian Internet research agency information operations in online conversations. *Defence Strategic Communications*, *6*, 155–198.

Gehlback, H., Robinson, C. D., & Vriesema, C. C. (2018). Climate conversations: Seeking a common starting point. *PsyArXiv*. https://doi.org/10.31234/osf.io/s8a7z

Gideon, L., Iii, C., Conway, K. R., & Houck, S. C. (2014). Automated integrative complexity. *Political Psychology*, *35*(5), 603–624. https://doi.org/10.1111/pops.12021

Google Project Jigsaw. (2018). *Perspective*. https://www.perspectiveapi.com/#/

Graf, S., Paolini, S., & Rubin, M. (2014). Negative intergroup contact is more influential, but positive intergroup contact is more common: Assessing contact prominence and contact prevalence in five Central European countries. *European Journal of Social Psychology*, *44*(6), 536–547. https://doi.org/10.1002/ejsp.2052

Gruzd, A., & Tsyganova, K. (2015). Information wars and online activism during the 2013/2014 crisis in Ukraine: Examining the social structures of Pro- and Anti-Maidan groups. *Policy and Internet*, *7*(2), 121–158. https://doi.org/10.1002/poi3.91

Guess, A., Barber, P., Vaccari, C., Kingdom, U., Nyhan, B., Seigel, A., . . .Stukal, D. (2018). *Social media, political polarization, and political disinformation: A review of the scientific literature*. William and Flora Hewlett Foundation. https://hewlett.org/library/social-media-political-polarization-political-disinformation-review-scientific-literature/

Guess, A., Lyons, B., Nyhan, B., & Reifler, J. (2018). *Avoiding the echo chamber about echo chambers: Why selective exposure to like-minded political news is less prevalent than you think*. Knight Foundation. https://kf-site-production.s3.amazonaws.com/media_elements/files/000/000/133/original/Topos_KF_White-Paper_Nyhan_V1.pdf

Guttieri, K., Wallace, M. D., & Suedfeld, P. (1995). The integrative complexity of American decision makers in the Cuban missile crisis. *Journal of Conflict Resolution*, *39*(4), 595–621. https://doi.org/10.1177/0022002795039004001

Hasler, B., & Amichai-Hamburger, Y. (2014). Online intergroup contact. In Y. Amichai-Hamburger (Ed.), *The social net: Understanding our online behavior* (2nd ed., pp. 220–252). Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199639540.003.0012

Herek, G. M., & Glunt, E. K. (1993). Interpersonal contact and heterosexuals' attitudes toward gay men: Results from a national survey. *The Journal of Sex Research*, *30*(3), 239–244. https://doi.org/10.1080/00224499309551707

Hoter, E. (2009). Information and Communication Technology (ICT) in the service of multiculturalism. *International Review of Research in Open and Distance Learning*, *10*(2), 1–15.

Houck, S. C. (2014). Automated integrative complexity : Current challenges and future directions. *Political Psychology*, *35*(5), 647–659. https://doi.org/10.1111/pops.12209

Houck, S. C., Repke, M. A., & Conway, L. G. (2017). Understanding what makes terrorist groups' propaganda effective: An integrative complexity analysis of ISIL and Al Qaeda. *Journal of Policing, Intelligence and Counter Terrorism*, *12*(2), 105–118. https://doi.org/10.1080/18335330.2017.1351032

Howard, P. N., Duffy, A., Freelon, D., Hussain, M. M., Mari, W., & Maziad, M. (2011). *Opening closed regimes: What was the role of social media during the Arab spring?* Project on Information Technology & Political Islam. https://doi.org/10.2139/ssrn.2595096

Ioannou, M., Hewstone, M., & Al Ramiah, A. (2017). Inducing similarities and differences in imagined contact: A mutual intergroup differentiation approach. *Group Processes and Intergroup Relations*, *20*(4), 427–446. https://doi.org/10.1177/1368430215612221

Islam, M. R., & Hewstone, M. (1993). Dimensions of contact as predictors of intergroup anxiety, perceived out-group variability, and out-group attitude: An integrative model. *Personality and Social Psychology Bulletin*, *19*(6), 700–710. https://doi.org/10.1177/0146167293196005

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly*, *76*(3), 405–431. https://doi.org/10.1093/poq/nfs038

Jackson, P. (2018). *The British extreme right: Reciprocal radicalisation and constructions of the other*. Radicalisation Research. https://www.radicalisationresearch.org/debate/jackson-british-extreme-right-reciprocal-radicalisation/

Jarvenpaa, S. L., & Leidner, D. E. (1999). Communication and trust in global virtual teams. *Organization Science*, *10*(6), 791–815. https://doi.org/10.1287/orsc.10.6.791

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical*

Methods in Natural Language Processing (pp. 1746–1751). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1181

Knott, K., Lee, B., & Copeland, S. (2018). *Briefings: Reciprocal radicalisation*. Centre for Research and Evidence on Security Threats. https://crestresearch.ac.uk/resources/reciprocal-radicalisation/

Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). *Community interaction and conflict on the Web*. ACM. https://doi.org/10.1145/3178876.3186141

Lee, E. J. (2007). Deindividuation effects on group polarization in computer-mediated communication: The role of group identification, public-self-awareness, and perceived argument quality. *Journal of Communication*, *57*(2), 385–403. https://doi.org/10.1111/j.1460-2466.2007.00348.x

MacInnis, C. C., & Page-Gould, E. (2015). How can intergroup interaction be bad if intergroup contact is good? Exploring and reconciling an apparent paradox in the science of intergroup relations. *Perspectives on Psychological Science*, *10*(3), 307–327. https://doi.org/10.1177/1745691614568482

Manbeck, K. E., Kanter, J. W., Kuczynski, A. M., Fine, L., Corey, M. D., & Maitland, D. W. M. (2018). Improving relations among conservatives and liberals on a college campus: A preliminary trial of a contextual-behavioral intervention. *Journal of Contextual Behavioral Science*, *10*, 120–125. https://doi.org/10.1016/j.jcbs.2018.10.006

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., . . .Mukherjee, A. (2019). Thou shalt not hate: Countering online hate speech. In *Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019, (ICWSM)* (pp. 369–380). https://www.aaai.org/ojs/index.php/ICWSM/article/download/3237/3105

Meleady, R., Seger, C. R., & Vermue, M. (2017). Examining the role of positive and negative intergroup contact and anti-immigrant prejudice in Brexit. *British Journal of Social Psychology*, *56*(4), 799–808. https://doi.org/10.1111/bjso.12203

Mitts, T. (2019). From isolation to radicalization: Anti-Muslim hostility and support for ISIS in the west. *American Political Science Review*, *113*(1), 173–194. https://doi.org/10.1017/S0003055418000618

Moghaddam, F. M. (2018). *Mutual radicalization: How groups and nations drive each other to extremes* (1st ed., *Vol. 163*). American Psychological Association. https://doi.org/10.1037/0000089-000

Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*, *2*(6), 389–396. https://doi.org/10.1038/s41562-018-0353-0

Mudde, C. (2007). *Populist radical right parties in Europe*. Cambridge University Press.

Mudde, C. (2019). *The far right today*. Polity.

Muhlhausen, D. B., & McNeill, J. B. (2011). *Terror trends: 40 years' data on international and domestic terrorism. Center for Data Analysis & Douglas and Sarah Allison Center for Foreign Policy Studies*. The Heritage Foundation. https://thf_media.s3.amazonaws.com/2011/pdf/sr0093.pdf

Müller, K., & Schwarz, C. (2018a). *Fanning the flames of hate: Social media and hate crime*. https://doi.org/10.2139/ssrn.3082972

Müller, K., & Schwarz, C. (2018b). *Making America hate again? Twitter and hate crime under Trump*. https://doi.org/10.2139/ssrn.3149103

Neiwert, D., Ankrom, D., Kaplan, E., & Pham, S. (2017). *Homegrown terrorism*. The Centre for Investigative Reporting. https://apps.revealnews.org/homegrown-terror/

O'Hara, K., & Stevens, D. (2015). Echo chambers and online radicalism: Assessing the Internet's complicity in violent extremism. *Policy and Internet*, *7*(4), 401–422. https://doi.org/10.1002/poi3.88

Omand, D., Bartlett, J., & Miller, C. (2012). Introducing social media intelligence (SOCMINT). *Intelligence and National Security*, *27*(6), 801–823. https://doi.org/10.1080/02684527.2012.716965

Paolini, S., Harwood, J., & Rubin, M. (2010). Negative intergroup contact makes group memberships salient: Explaining why intergroup conflict endures. *Personality and Social Psychology Bulletin*, *36*(12), 1723–1738. https://doi.org/10.1177/0146167210388667

Park, G., & DeShon, R. P. (2018). Effects of group-discussion integrative complexity on intergroup relations in a social dilemma. *Organizational Behavior and Human Decision Processes*, *146*, 62–75. https://doi.org/10.1016/j.obhdp.2018.04.001

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*(5), 751–783. https://doi.org/10.1037/0022-3514.90.5.751

Pettigrew, T. F., & Tropp, L. R. (2008). How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. *European Journal of Social Psychology*, *38*, 922–934. https://doi.org/10.1002/ejsp

Postmes, T., Spears, R., & Lea, M. (1998). Building or breaching social boundaries? SIDE effects of computer mediated communication. *Communication Research*, *25*(6), 689–715.

Poushter, J., Wike, R., & Oates, R. (2015). *Extremism concerns growing in west and predominantly Muslim countries*. Pew Research Centre. https://www.pewresearch.org/global/2015/07/16/extremism-concerns-growing-in-west-and-predominantly-muslim-countries/

Pratt, D. (2015). Islamophobia as reactive co-radicalization. *Islam and Christian–Muslim Relations*, *26*(2), 205–218. https://doi.org/10.1080/09596410.2014.1000025

Ramiah, A., & Al Hewstone, M. (2013). Intergroup contact as a tool for reducing, resolving, and preventing intergroup conflict: Evidence, limitations, and potential. *American Psychologist*, *68*(7), 527–542. https://doi.org/10.1037/a0032603

Ruesch, M. (2011). A peaceful Net? In First Global Conference on Communication and Conflict (pp. 1–19). https://www.lse.ac.uk/media-and-communications/assets/documents/alumni/Michelle-Ruesch-CCConference-forLSE-final.pdf

Schroeder, J., Kardas, M., & Epley, N. (2017). The humanizing voice: Speech reveals, and text conceals, a more thoughtful mind in the midst of disagreement. *Psychological Science*, *28*(12), 1745–1762. https://doi.org/10.1177/0956797617713798

Schumann, S., Klein, O., Douglas, K., & Hewstone, M. (2017). When is computer-mediated intergroup contact most promising? Examining the effect of out-group members' anonymity on prejudice. *Computers in Human Behavior*, *77*, 198–210. https://doi.org/10.1016/j.chb.2017.08.006

Schwab, A. K., & Greitemeyer, T. (2015). The world's biggest salad bowl: Facebook connecting cultures. *Journal of Applied Social Psychology*, *45*(4), 243–252. https://doi.org/10.1111/jasp.12291

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. https://doi.org/10.1145/505282.505283

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly Media.

Sirseloudi, M. (2017, April). Dyadic radicalisation via internet propaganda [Conference session]. Europol Conference on Online Terrorist Propaganda, The Hague, The Netherlands.

Smith, A., Suedfeld, P., Conway, L., & Winter, D. (2008). The language of violence: Distinguishing terrorist from nonterrorist groups by thematic content analysis. *Dynamics of Asymmetric Conflict*, *1*(2), 142–163. https://doi.org/10.1080/17467580802590449

Sønderskov, K. M., & Thomsen, J. P. F. (2015). Contextualizing intergroup contact: Do political party cues enhance contact effects? *Social Psychology Quarterly*, *78*(1), 49–76. https://doi.org/10.1177/0190272514560761

Steinert-Threlkeld, Z. C., Mocanu, D., Vespignani, A., & Fowler, J. (2015). Online social networks and offline protest. *EPJ Data Science*, *4*(1), 1–9. https://doi.org/10.1140/epjds/s13688-015-0056-y

Streufert, S., & Suedfeld, P. (1965). Conceptual structure, information search, and information utilization. *Journal of Personality and Social Psychology*, 2(5), 736–740. http://www.ncbi.nlm.nih.gov/pubmed/5838772

Suedfeld, P. (2010). The cognitive processing of politics and politicians: Archival studies of conceptual and integrative complexity. *Journal of Personality*, *78*(6), 1669–1702. https://doi.org/10.1111/j.1467-6494.2010.00666.x

Suedfeld, P., & Bluck, S. (1988). Changes in integrative complexity prior to surprise attacks. *Journal of Conflict Resolution*, *32*(4), 626–635. https://doi.org/10.1177/0022002788032004002

Suhay, E., Bello-Pardo, E., & Maurer, B. (2018). The polarizing effects of online partisan criticism: Evidence from two experiments. *International Journal of Press/Politics*, *23*(1), 95–115. https://doi.org/10.1177/1940161217740697

Sunstein, C. R. (2017). *#Republic: Divided democracy in the age of social media*. Princeton University Press.

Tetlock, P. E., Peterson, R. S., & Berry, J. M. (1993). Flattering and unflattering personality portraits of integratively simple and complex managers. *Journal of Personality and Social Psychology*, *64*(3), 500–511. https://doi.org/10.1037/0022-3514.64.3.500

Tufekci, Z. (2017). *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.

Von Behr, I., Reding, A., Edwards, C., & Gribbon, L. (2013). *Radicalisation in the digital era: The use of the internet in 15 cases of terrorism and extremism*. RAND. https://www.rand.org/pubs/research_reports/RR453.html

Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal and hyperpersonal interaction. *Communication Research*, *23*(1), 3–43.

White, F. A., & Abu-Rayya, H. M. (2012). A dual identity-electronic contact (DIEC) experiment promoting short- and long-term intergroup harmony. *Journal of Experimental Social Psychology*, *48*(3), 597–608. https://doi.org/10.1016/j.jesp.2012.01.007

Wildschut, T., Pinter, B., Vevea, J. L., Insko, C. A., & Schopler, J. (2003). Beyond the group mind: A quantitative review of the interindividual-intergroup discontinuity effect. *Psychological Bulletin*, *129*(5), 698–722. https://doi.org/10.1037/0033-2909.129.5.698

Williams, H. T. P., McMurray, J. R., Kurz, T., & Hugo-Lambert, F. (2015). Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change*, *32*, 126–138. https://doi.org/10.1016/j.gloenvcha.2015.03.006

Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2019). Hate in the machine: Anti-black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, *60*, 93–117. https://doi.org/10.1093/bjc/azz049

Wołk, K., & Marasek, K. (2015). Neural-based machine translation for medical text domain. Based on European Medicines Agency leaflet texts. *Procedia Computer Science*, *64*, 2–9. https://doi.org/10.1016/j.procs.2015.08.456

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina: Personal attacks seen at scale. In *International World Wide Web Conference* (pp. 1391–1399). ACM. https://doi.org/10.1145/3038912.3052591

Yardi, S., & Boyd, D. (2010). Dynamic debates: An analysis of group polarization over time on Twitter. *Bulletin of Science, Technology & Society*, *30*(5), 316–327. https://doi.org/10.1177/0270467610380011

Author Biographies

John D. Gallacher is a DPhil student at the University of Oxford, working in the Cyber Security Center for Doctoral Training, the Oxford Internet Institute and the Department for Experimental Psychology. His research focuses on the dynamics of online communication and their implications on extremism, radicalization and intergroup conflict.

Marc W. Heerdink is an assistant professor of Social Psychology at the University of Amsterdam, where he also obtained his PhD, and a former visiting postdoc at the University of Oxford. His research focuses on understanding the interplay between emotional behavior and group dynamics.

Miles Hewstone is an emeritus professor at the University of Oxford, with research interests in the field of experimental social psychology, focusing on prejudice and stereotyping, intergroup contact and the reduction of intergroup conflict.