Countering Extremists on Social Media

Ganesh, Bharath; Bright, Jonathan

Published in:
Policy & Internet

DOI:
10.1002/poi3.236

Publication date:
2020

Countering extremists on social media: challenges for strategic communication and content moderation

Guest Editorial for *Policy & Internet*

Bharath Ganesh[1,2] and Jonathan Bright[1]

[1] Oxford Internet Institute, University of Oxford
[2] Centre for Media and Journalism Studies, University of Groningen

Word count: 3,897 (5,897 including references)

Please consult the publisher's version of record:
Ganesh, B. and Bright, J. (2020). Countering extremists on social media: challenges for strategic communication and content moderation. *Policy & Internet* 12(1): 6-19. DOI: 10.1002/poi3.236.

Introduction

Extremist exploitation of social media platforms is an important regulatory question for civil society, government, and the private sector (Crosset and Dupont 2018), mirroring existing discussions about platform governance in general (Gorwa 2019). Extremists exploit social media platforms, and the Internet more generally, for a range of reasons from spreading hateful narratives and propaganda to financing, recruitment, and sharing operational information (Gill *et al.* 2017). How best to counter such activity has recently been the focus of an emerging field of academic and policy debate (Aly *et al.* 2016, Braddock and Horgan 2016, Davies *et al.* 2016, Szmania and Fincher 2017, Helmus 2018, Ganesh and Bright 2020). While many extremists end up barred from social media at the discretion of hosting platforms (Citron 2018, Gillespie 2018), often in discussion with government and law enforcement (Brocato 2015, Brown and Pearson 2018), significant attention is being paid to counter-messaging and other strategic communication techniques as potential responses (Briggs and Feve 2013, Bertram 2016, Beutel *et al.* 2016, Braddock and Horgan 2016, Cherney 2016, Brown and Marway 2018, Eerten *et al.* 2019). How best to respond to extremism on social media often centres on the vexing task of finding a balance between civil society, government, and private sector actors and a balance in regulating and moderating content on platforms and developing programmes to counter the narratives on which extremists thrive while being conscious of rights to free expression and the appropriateness of restrictions on speech.

Policy responses to this question fit under two headings: *strategic communication* and *content moderation*. This issue focuses on one form of strategic communication, countering violent extremism (CVE) which we introduce in the following section (see Archetti, 2019). Content moderation, which is different than CVE though it affects extremist exploitation of social media, is a set of practices used by social media platforms to enforce their guidelines on acceptable content. As we describe below, there are emerging relationships across civil society,

government, and private sector actors in content moderation. At the centre of both of these policy responses is a calculation about how best to limit audience exposure to extremist narratives and maintain the marginality of extremist views. Extremists, meanwhile, seek to use social media to expand their reach, appear credible, and transgress this marginality. Challenging extremists on social media requires a variety of techniques, and increasingly relies on groups of stakeholders across civil society, and the private sector, rather than just government alone (Briggs and Feve 2013, Griffith-Dickson *et al.* 2014, Aly *et al.* 2015, Dalgaard-Nielsen 2016, Scrivens and Perry 2017, Brown and Marway 2018, Gielen 2019). Strategic communication and content moderation are two broad responses to consider in policy development to challenge extremist exploitation of social media.

This issue collects five articles that develop multiple strands of research into the responses and solutions to extremist exploitation of social media. Through these five articles, we suggest an agenda for future research on how multi-stakeholder initiatives to challenge extremist exploitation of social media are conceived, designed, and implemented, and what challenges these initiatives need to surmount.

Strategic Communication, Primary CVE, and Informal Actors

Countering violent extremism (CVE) refers to a field of "soft power" mechanisms that try to counter extremists, and should be differentiated from counter-terrorism. CVE seeks to use "non-coercive" and "voluntary" activities designed to counter violent extremist ideology and attempts to provide opportunities for individuals to disengage with radicalising influences (Selim 2016, p. 95, Bjola and Pamment 2019, p. 7). Alongside working with local communities and supporting individuals, strategic communications is one of the key functions of CVE.[1]

[1] The US-led Global Engagement Centre, the UK Home Office's RICU (Research, Information and Communications Unit), and the independent Hedayah Centre based in Abu Dhabi are well-known centres for CVE knowledge and practice, all of which ground CVE activity in strategic communications (Archetti 2019, pp. 85–86).

Many CVE programmes are funded by governments, but often delivered by civil society, such as the EU's Civil Society Empowerment Programme, or the private sector as is the case in the UK (described below). Broadly, CVE initiatives incorporate contributions from civil society, governments, think tanks and non-profits, and the private sector.

CVE activities can be conceptualised as primary, secondary, and tertiary. Primary CVE seeks to reduce the likelihood of radicalisation across a population, secondary CVE focuses on those vulnerable to radicalisation, and tertiary CVE focuses on those already radicalised (Harris-Hogan *et al.* 2016, Gielen 2019, p. 1157). While secondary and tertiary CVE often involve state-run exit and deradicalization programmes, frequently making use of civil society practitioners and social workers, primary CVE focuses on challenging the spread of extremist narratives and inoculating audiences against them. The articles collected in this special issue offer new avenues for conceiving of primary CVE activities on and through social media, and explore how they can be refined by learning from previous CVE initiatives, informal CVE actors, and organic activity on platforms.

In addition to civil society, think tanks, and the government, the private sector now plays an important role in primary CVE. First, social media platforms that extremists exploit have become key stakeholders in the governance of extremism. This means that Facebook, Twitter, and Alphabet/Google have become important actors in countering extremism on the platforms they run. For example, Facebook has developed in-house technologies and protocols,[2] is working with civil society for counter-messaging and anti-hate work,[3] and moderates content and suspends users where necessary.[4] Second, the cultural industries—particularly advertising, public relations, and media production—have been contracted by the state to produce counter-narrative content. A well-known example is the UK Home Office's

[2] https://about.fb.com/news/2017/06/how-we-counter-terrorism/
[3] https://www.lifeafterhate.org/blog/2019/3/27/life-after-hate-working-with-facebook-to-help-individuals-leave-behind-hate-groups
[4] https://transparency.facebook.com/community-standards-enforcement#instagram-terrorist-propaganda

RICU contracting Breakthrough Media,[5] a production company, to produce content that challenges violent jihadist narratives. The UK's Home Office also contracted M&C Saatchi, a major advertising company, to manage a GBP 60 million account to develop CVE campaigns,[6] which has continued in the UK under the "Building a Stronger Britain Together" programme run by the Home Office.[7] While it remains to be seen what effect this investment has had on preventing extremism and disrupting circuits of radicalisation, this is clear evidence that more stakeholders in the cultural industries are increasingly becoming involved in governance processes to counter extremist exploitation of digital media. Much of this work proceeds without significant academic scrutiny and evaluation, often with thin evidence that these initiatives are indeed as effective as they promise to be (Glazzard 2017, Archetti 2019, Awan *et al.* 2019). When used in conjunction with automated recommendation systems, they may even risk counter-productive effects (Schmitt *et al.* 2018, Bright et al. 2020).

Of course, CVE has not primarily been focused on online initiatives, though many CVE service providers have recently increased their attention to extremist exploitation of social media. The first article in this special issue by Talene Bilazarian studies three cases of formal, offline CVE initiatives led by the state and third parties. Bilazarian argues that overt participation from a government may compromise the credibility of CVE activity (see also Neumann 2013, Ingram 2016, Belanger and Szmania 2018). She suggests that messages from third parties can alleviate these concerns about credibility. Third parties, she argues, are better placed to take advantage of existing network effects and use interactive features to increase the impact of CVE efforts. Bilazarian's recommendation to focus on networked approaches, interpersonal messaging, and going beyond the narrow frame of counter-extremism when considering relevant actors in online CVE sets the stage for Lee's work on informal counter-

[5] https://www.theguardian.com/politics/2016/may/02/inside-ricu-the-shadowy-propaganda-unit-inspired-by-the-cold-war
[6] https://www.theguardian.com/world/2017/feb/06/uks-government-hires-advertising-giant-as-it-fights-far-right-threat
[7] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/836780/building-stronger-britain-together-2019-horr112.pdf

narratives and Chaudhry and Gruzd's work in comment section racism on Facebook news pages. Where Bilazarian develops policy recommendations that can better guide online-oriented CVE, Lee and Chaudhry and Gruzd provide a granular examination of the challenges facing primary CVE on and through social media.

Given the increased participation of civil society, the private sector, and the cultural industries in CVE, Lee asks, "why would audiences listen to a word the counter messaging 'industry' has to say?" (p. 2). Lee's article shifts focus to informal counter messaging, understood as "spontaneous", everyday expressions that are "inherent in societies" that "maintain the social prohibition on extreme ideas and behaviours" (p. 5-6). Such users are important to CVE efforts because they present independent, and possibly more 'credible' voices for counter-messaging (Coyer 2020).

Turning to the experiences of informal counter-narrative practitioners, Lee concludes that their focus on satirising, criticising, and challenging extremist narratives contributes to primary CVE by reinforcing social prohibitions against such views in mainstream venues. Indeed, these informal mechanisms are more and more part of formal strategic communications. By identifying key challenges, particularly around ideology, motivation, and shared values, Lee reveals some of the challenges that will be faced in the future as relatively powerful actors continue employ civil society to take part in strategic communications intended to disrupt extremist use of digital media.

Informal ensembles of users also play a role in primary CVE, though they cannot be classified as engaging in strategic communication. Rather, we can look to users on platforms as another set of informal actors challenging extremist narratives. Drawing on empirical research on thousands of comments on news stories about race, racism, or ethnicity on the Canadian Broadcasting Corporation News Facebook page, Chaudhry and Gruzd focus on the "spiral of silence" which is a communication theory that "suggests that with increasing social

pressure, people may conceal their views when they think their views are in the minority" (see Noelle-Neumann 1991 in Chaudhry and Gruzd, p. 2). Though they suggest that the lack of anonymity on Facebook limits the extent of racist speech observed on the page they study, Chaudhry and Gruzd do find a vocal minority of users participating in racist speech. However, they also find that a sizable proportion of users take it upon themselves to counter racist narratives when they are expressed by other users on the page. This work expands on the possibilities and limits of ensembles of users participating in forms of primary CVE in an organic and self-directed fashion that is not typically associated with CVE efforts, providing crucial data on the possibilities and limits of incorporating such actors in efforts to engage in primary CVE.

Content Moderation and Takedown

Content moderation references another set of policy responses that are not forms of strategic communication or CVE. However, content moderation has an effect in the same fields in which primary CVE intervenes because content moderation involves decisions about decreasing the presence of extremist narratives or suspending exponents of extremist on a platform, thereby reducing the potential that audiences might be exposed to extremist narratives. Content moderation is done by social media platforms, who use large labour forces, often with acute effects on the mental health of precarious workers, and automated tools to identify extremist content, defined by each platforms' own community guidelines (Gillespie 2018, Roberts 2019). Platforms are in charge of enforcing these guidelines and regularly remove content and block users that are in violation of guidelines that they have set on hate speech, inappropriate content, support or celebration of terrorism, or spam.

This is a controversial area, but Conway et al. find that Twitter takedown of pro-IS accounts "severely affected IS's ability to develop and maintain robust and influential

communities on Twitter" (Berger and Perez 2016, Conway *et al.* 2019, p. 152). On Reddit, users active on hate-based subforums that were shut down became active on other parts of Reddit, but their expression of hate, misogyny, and racism had decreased (Chandrasekharan *et al.* 2017). While taking down extremism may seem a logical approach, it can have counter-productive outcomes. First, disruption on Twitter has led to the migration of pro-ISIS activity to encrypted messaging applications such as Telegram (Prucha 2016). Second, having faced suspension on Twitter and other social media platforms can be a badge of pride for extremists and plays a role in community-building among these networks (Pearson 2017).

Content moderation also involves multiple stakeholders that include government (particularly law enforcement) and civil society. For example, Internet Referral Units run by police organizations such as Europol and London's Metropolitan Police, play an important role in encouraging platforms to take down content (Chang 2017, Vieth 2019, Reeve 2020). Further, social media companies have developed their own relationship with specific civil society organisations that it has selected as 'trusted flaggers' of potentially extremist content (Fishman 2019, p. 93). The Global Internet Forum to Counter Terrorism (GIFCT) is a further development in this area, which involves a shared database of image fingerprints (or "hashes") to enable rapid takedown of extremist content across platforms and websites. It also brings together multiple stakeholders and works with the UN, intergovernmental organisations, think tanks, and civil society (Gorwa 2019). More recently, there has been efforts across computer science and computational linguistics specialists in the academy and industry to develop reliable systems that can detect extremist expression on social media, using text mining, classification, and image recognition techniques (Djuric *et al.* 2015, Burnap and Williams 2016, Rudinac *et al.* 2017, Borisyuk *et al.* 2018, Scrivens *et al.* 2018). The initiatives are occurring alongside the increase in private sector initiatives that use AI to detect and assist with moderation of extremist content (Gallacher 2020).

Research on emerging technologies in moderating extremist and terrorist content requires more attention. Given the high risks of incorrect flags that lead to takedown of innocent users and their content, auditing and evaluating AI approaches at use in content moderation is of significant concern, especially considering the demonstrable biases against women and minorities that studies of algorithms have revealed (Eubanks 2018, Noble 2018). While many projects have focused on how to detect extremist content, Hall et al. instead evaluate the performance of machines against human judgement, probing the limits of text-based methods for the classification of extremism. They find that for jihadist content, approaches to detect extremist content with AI require significant work in integrating human understanding into machine abilities. While these approaches perform well for high-level concepts, humans provide more granular analysis that identifies key themes and forms of content, such as emotion. By engaging in a validation of open-source AI tools in detection of extremist content, Hall et al. provide valuable advancements in research design and methodology that can be applied for future study that probes the possibilities and limits of technical systems in primary CVE, identifying key challenges that software must surmount for it to be a viable alternative to human-led moderation.

Alternative Media Undermining Primary CVE

While CVE practitioners are acutely aware of the broader networks of websites and blogs that form an alternative media ecosystem that provides an important resource for extremists, this ecosystem presents significant challenges in ensuring the viability of primary CVE activities. Research on disinformation, so-called "fake news", and polarization on social media has highlighted the important roles played by users in spreading fake news (Vosoughi *et al.* 2018),

the unlikelihood of extremists engaging with others that represent rival ideologies (Bright 2018), the higher likelihood of conservatives sharing stories from fake news domains (Guess *et al.* 2019), and the disproportionate role of radical right media in the spread of disinformation (Bennett and Livingston 2018). Recent work has stressed the central role that alternative media—such as Breitbart News and the social media use of far right social movements in North America and Western Europe—has had in spreading problematic information and reinforcing discriminatory, racist discourse and positions common to both the radical right and the extreme right (Benkler *et al.* 2018, Bennett and Livingston 2018, Marwick 2018).

In exploiting social media, extremists are taking advantage of communication infrastructure, the specific affordances and cultures specific to certain platforms, media gatekeepers, and multiple networked audiences to whom they deliver content. Alternative media can be anti-democratic, repressive, and denigrating to out-groups while continuing to challenge hegemonic discourse in mainstream media (see p. 4 of Heft *et al*. in this issue). If a social prohibition on extremist views—jihadist, far right, or otherwise—is important to uphold, primary CVE should consider the extent to which alternative media networks exploit infrastructure, affordances, and different media systems to undermine this prohibition.

The role of civil society, government, and social media platforms in addressing issues of disinformation and hyper-partisan alternative media is central to primary CVE; consequently, an engagement (offered by Heft et al. in this issue) is necessary to understanding what Marwick (2018) refers to as the sociotechnical systems in which racist, discriminatory, and hateful disinformation is mediated, potentially undermining primary CVE by legitimising and reinforcing forms of anti-immigrant, racist, and white supremacist discourse through alternative media networks between users, political actors, and a variety of creators active on platforms that are connected with participatory digital cultures on social media platforms (Lewis 2018, Marwick 2018, Hughes 2019, Munn 2019).

While public attention has shifted to ISIS media in recent years, jihadists have long developed news outlets, webforums, and networks to share training and technical manuals (Awan 2007, pp. 76–77, Hoskins *et al.* 2011, Archetti 2012, Conway *et al.* 2012). This alternative media network was part of a strategy to "break the media siege imposed on the jihad movement" (Ayman Al-Zawahiri in Awan 2007, p. 76). They developed techniques to enhance their legitimacy and appear as credible websites while exploiting cynicism and mistrust with Western news sources amongst audiences in the UK (2007, 78).

In the past decade, production values of jihadist alternative media have increased considerably and were widely disseminated on social media platforms, helping to maintain their presence online (Fisher 2015, Al-Rawi 2018, Shehabat and Mitew 2018, Baele *et al.* 2019, Fisher *et al.* 2019, Winter 2019). However, recent efforts at platform governance involving governments, civil society, social media platforms and internet companies have had a significant effect on forcing jihadists off major platforms and applications (Conway *et al.* 2019). While it is not clear that this has fully countered their ability to maintain a persistent jihadist alternative media network available on the surface web, it has made mainstream platforms less accessible to them. Moreover, platforms such as Facebook, Twitter, and YouTube work with smaller platforms and other service providers to share data about jihadist content and automate pre-emptive content moderation, and more attention is expected to be paid to the extreme right, especially following the Christchurch Call to Action.[8]

Jihadist alternative media have a very different relationship to platform governance than right-wing alternative media. Far right narratives are readily accessible on social media platforms and their exponents and audiences often benefit from legitimation from political representatives in Western democracies (see Benkler *et al.* 2018, Ch. 3). Platforms such as YouTube facilitate forms of microcelebrity and interconnectivity between extremist content

[8] https://gifct.org/press/global-internet-forum-counter-terrorism-update-our-progress-two-years/

creators that confer legitimacy and credibility on these creators (Lewis 2020). This makes it much more difficult for a set of actors to act decisively to counter such content—if these narratives are repeated by elected representatives, how should social media platforms react to such content? As noted by Twitter employees in a recent article published by *Vice News*, targeting American white supremacists on the platform may also involve banning Republican politicians.[9] Takedown efforts have been met with a significant backlash and so-called "alt-tech" platforms have become a home for extremists banned from mainstream platforms, providing a relatively secure site for extremist narratives to circulate (Donovan *et al.* 2018).

As the authors of the final contribution to this special issue note, alternative media on the political right "results in a combination of an anti-hegemonic impetus" and a wide range of both mainstream and extreme political positions, ranging from economic liberalism to nativism (p. 5). While research on far right exploitation of social media is increasing, much of this work focuses on representation, narratives, ideology, and discourse (Topinka 2018, Deem 2019, Froio and Ganesh 2019, Klein and Muis 2019, Richards 2019), as well as disinformation spreading from right-wing digital news to social media platforms and mainstream media (Marwick and Lewis 2017, Benkler *et al.* 2018, Bennett and Livingston 2018). Developing the context of such activity in comparative perspective, Heft et al. in this volume provide a thorough mapping of hyper-partisan outlets in right-wing digital news ecosystems in Austria, Denmark, Germany, Sweden, the United Kingdom, and the United States that contributes a context for both of these lines of inquiry. The authors identify contextual factors in each country's political and media system that has led to different configurations of right-wing alternative media online in each country, findings various structures, styles, and supply and demand markets.

[9] https://www.vice.com/en_us/article/a3xgq5/why-wont-twitter-treat-white-supremacy-like-isis-because-it-would-mean-banning-some-republican-politicians-too

The authors classify 70 websites in this alternative media system by various factors including their tendency (a measurement of how conventionally their site is structured versus how focused their site is on sensational right-wing topics), transparency, and advertising dependency. Heft et al. ultimately note that they find "different patterns of supply and demand, as well as distinct funding structures, organizational strategies, and thematic tendency" across all of the sites (p. 24). More importantly, they find that right-wing digital news is tending towards normalization, which "challenges digital news environments" because normalization makes it more difficult for audiences to differentiate hyper-partisan from regular news (p. 24). However, they also note the significance of transnational audiences; while there is significant heterogeneity in the news pages they explored, English-based right-wing digital news enjoys transnational audiences.

While the media Heft et al. explore cannot be uniformly or uncontroversially referred to as extreme right, nor is it directly implicated in far right terrorism, they do demonstrate a number of significant trends relevant to the development of solutions to counter extremist exploitation of social media. By repeating nativist, xenophobic, anti-Muslim, anti-Semitic, and anti-establishment themes, right-wing alternative news create an environment in which non-violent extremist subcultures can thrive (see Holt *et al.* 2017). While there is little research to prove that these non-violent extremist subcultures *cause* violence, it is clear that they provide a milieu in which extremist views are sanctioned, supported, and reinforced rather than challenged and marginalized. Thus, alternative news can media undermine efforts at primary CVE and must be understood as actors that present a challenge to both formal and informal counter-messaging. The last article in this special issue contributes an overview of challenges faced by attempts at governing extremist exploitation of social media and the key role that alternative media play in supporting and cultivating a milieu that degrades the social prohibitions against ring-wing extremist views.

Conclusion

Research into extremist exploitation of social media is a rapidly-developing field, as is research into the design, development, and implementation of counter-measures. In this introduction, we have introduced the contested role between civil society, government, and the private sector in initiatives to counter extremist exploitation of social media. We argue that these three actors play an important role in primary CVE, particularly in terms of strategic communication and content moderation. Across our articles focused on strategic communication, we see that emphasis is placed on the potential of informal actors to challenge and reinforce social norms that reject extremist views. Turning to content moderation, a more blunt tool that *enforces* these norms, we find that the increased involvement of new technologies requires auditing and criticism to identify the reliability of automation in such a contentious, high-risk area. Finally, looking at the potential of alternative media to chip away at these social injunctions against extremism, we explore how the mapping of the right-wing alternative media across different countries reveals significant heterogeneity and the processes by which extremist views are normalized through alternative media. The five articles collected in this issue provide an initial foray into encouraging interdisciplinary research on the challenges, possibilities, and limits of tools in use to counter extremist exploitation of social media.

References

Al-Rawi, A., 2018. Video games, terrorism, and ISIS's Jihad 3.0. *Terrorism and Political Violence*, 30 (4), 740–760.

Aly, A., Balbi, A.-M., and Jacques, C., 2015. Rethinking countering violent extremism: implementing the role of civil society. *Journal of Policing, Intelligence and Counter Terrorism*, 10 (1), 3–13.

Aly, A., Macdonald, S., Jarvis, L., and Chen, T., 2016. *Violent Extremism Online: New Perspectives on Terrorism and the Internet*. Routledge.

Archetti, C., 2012. *Understanding Terrorism in the Age of Global Media: A Communication Approach*. Palgrave Macmillan.

Archetti, C., 2019. The unbearable thinness of strategic communication. *In*: C. Bjola and J. Pamment, eds. *Countering Online Propaganda and Extremism: The Dark Side of Digital Diplomacy*. Routledge, 81–96.

Awan, A., Miskimmon, A., and O'Loughlin, B., 2019. The battle for the battle of the narratives: side-stepping the double fetish of digital and CVE. *In*: C. Bjola and J. Pamment, eds. *Countering Online Propaganda and Extremism: The Dark Side of Digital Diplomacy*. Routledge, 156–171.

Awan, A.N., 2007. Radicalization on the Internet? *The RUSI Journal*, 152 (3), 76–81.

Baele, S.J., Boyd, K., and Coan, T., 2019. *ISIS Propaganda: A Full-Spectrum Extremist Message*. Oxford: Oxford University Press.

Belanger, P. and Szmania, S., 2018. The Paradox of Source Credibility in Canadian and U.S. Domestic Counterterrorism Communications. *International Journal of Communication*, 12 (0), 22.

Benkler, Y., Faris, R., and Roberts, H., 2018. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford, UK: Oxford University Press.

Bennett, W.L. and Livingston, S., 2018. The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, 33 (2), 122–139.

Berger, J.M. and Perez, H., 2016. *The Islamic State's Diminishing Returns on Twitter: How Suspensions are Limiting the Social Networks of English-speaking ISIS Supporters*. George Washington University.

Bertram, L., 2016. Terrorism, the Internet and the Social Media Advantage: Exploring how terrorist organizations exploit aspects of the internet, social media and how these same platforms could be used to counter-violent extremism. *Journal for Deradicalization*, 0 (7), 225–252.

Beutel, A., Weine, S., Saeed, A., Mihajlovic, A., Stone, A., Beahrs, J., and Shanfield, S., 2016. Guiding Principles for Countering and Displacing Extremist Narratives. *Contemporary Voices: St Andrews Journal of International Relations*, 7 (3).

Bilazarian, T. 2020. Countering Violent Extremist Narratives Online: Lessons From Offline Countering Violent Extremism. *Policy & Internet* 12(1): 46–65.

Bjola, C. and Pamment, J., 2019. Introduction: the 'dark side' of digital diplomacy. *In*: C. Bjola and J. Pamment, eds. *Countering Online Propaganda and Extremism: The Dark Side of Digital Diplomacy*. Routledge, 1–10.

Borisyuk, F., Gordo, A., and Sivakumar, V., 2018. Rosetta: Large Scale System for Text Detection and Recognition in Images. *In*: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18*. Presented at the the 24th ACM SIGKDD International Conference, London, United Kingdom: ACM Press, 71–79.

Braddock, K. and Horgan, J., 2016. Towards a Guide for Constructing and Disseminating Counternarratives to Reduce Support for Terrorism. *Studies in Conflict & Terrorism*, 39 (5), 381–404.

Briggs, R. and Feve, S., 2013. *Review of Programs to Counter Narratives of Violent Extremism*. London: Institute for Strategic Dialogue.

Bright, J., 2018. Explaining the Emergence of Political Fragmentation on Social Media: The Role of Ideology and Extremism. *Journal of Computer-Mediated Communication*, 23 (1), 17–33.

Brocato, A., 2015. Tackling Terrorists' Use of the Internet: Propaganda Dispersion & the Threat of Radicalization. *In*: M.N. Ogun, ed. *Terrorist Use of Cyberspace and Cyber Terrorism: New Challenges and Responses*. IOS Press, 129–148.

Brown, K. and Marway, H., 2018. *Preventing Radicalisation to Terrorism and Violent Extremism: Delivering counter- or alternative narratives*. Radicalisation Awareness Network.

Brown, K. and Pearson, E., 2018. Social Media, The Online Environment and Terrorism. *In*: A. Silke, ed. *Routledge Handbook of Terrorism and Counterterrorism*. London: Routledge.

Burnap, P. and Williams, M.L., 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, 5 (1), 11.

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., and Gilbert, E., 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proc. ACM Hum.-Comput. Interact.*, 1 (CSCW), 31:1–31:22.

Chang, B., 2017. From Internet Referral Units to International Agreements; Censorship of the Internet by the UK and EU. *Columbia Human Rights Law Review*, 49, 114.

Chaudhry, I., and A. Gruzd. 2020. Expressing and Challenging Racist Discourse on Facebook: How Social Media Weaken the "Spiral of Silence" Theory. *Policy & Internet* 12(1): 88–108

Cherney, A., 2016. Designing and implementing programmes to tackle radicalization and violent extremism: lessons from criminology. *Dynamics of Asymmetric Conflict*, 9 (1–3), 82–94.

Citron, D.K., 2018. *Extremist Speech, Compelled Conformity, and Censorship Creep*. Rochester, NY: Social Science Research Network, SSRN Scholarly Paper No. ID 2941880.

Conway, M., Khawaja, M., Lakhani, S., Reffin, J., Robertson, A., and Weir, D., 2019. Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts. *Studies in Conflict & Terrorism*, 42 (1–2), 141–160.

Conway, M., McInerney, L., and Ducol, B., 2012. Uncovering the French-speaking jihadisphere: An exploratory analysis. *Media, War & Conflict*, 5 (1), 51–70.

Coyer, K., 2020. Informal Counter-Narratives. *In*: B. Ganesh and J. Bright, eds. *Extreme Digital Speech: Contexts, Responses, and Solutions*. Dublin: VOX-Pol.

Crosset, V. and Dupont, B., 2018. Internet et propagande jihadiste : la régulation polycentrique du cyberespace. *Critique internationale*, N° 78 (1), 107–125.

Dalgaard-Nielsen, A., 2016. Countering Violent Extremism with Governance Networks. *Perspectives on Terrorism*, 10 (6).

Davies, G., Neudecker, C., Ouellet, M., Bouchard, M., and Ducol, B., 2016. Toward a Framework Understanding of Online Programs for Countering Violent Extremism. *Journal for Deradicalization*, 0 (6), 51–86.

Deem, A., 2019. The Digital Traces of #whitegenocide and Alt-Right Affective Economies of Transgression. *International Journal of Communication*, 13 (0), 20.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., and Bhamidipati, N., 2015. Hate Speech Detection with Comment Embeddings. *In*: *Proceedings of the 24th International Conference on World Wide Web*. New York, NY, USA: ACM, 29–30.

Donovan, J., Lewis, B., and Friedberg, B., 2018. Parallel Ports: Sociotechnical Change from the Alt-Right to Alt-Tech. *In*: M. Fielitz and N. Thurston, eds. *Post-Digital Cultures of the Far Right: Online Actions and Offline Consequences in Europe and the US*. transcript Verlag.

Eerten, J.-J. van, Doosje, B., and Doosje, B., 2019. *Challenging Extremist Views on Social Media : Developing a Counter-Messaging Response*. Routledge.

Eubanks, V., 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Publishing Group.

Fisher, A., 2015. How Jihadist Networks Maintain a Persistent Online Presence. *Perspectives on Terrorism*, 9 (3).

Fisher, A., Prucha, N., and Winterbotham, 2019. *Mapping the Jihadist Information System: Towards the Next Generation of Disruption Capability*. London: Royal United Services Institute.

Fishman, B., 2019. Crossroads: Counter-Terrorism and the Internet. *Texas National Security Review*, 2 (2), 82–100.

Froio, C. and Ganesh, B., 2019. The transnationalisation of far right discourse on Twitter. *European Societies*, 21 (4), 513–539.

Gallacher, J., 2020. Automated detection of terrorist and extremist content. *In*: B. Ganesh and J. Bright, eds. *Extreme Digital Speech: Contexts, Responses, and Solutions*. Dublin: VOX-Pol.

Gielen, A.-J., 2019. Countering Violent Extremism: A Realist Review for Assessing What Works, for Whom, in What Circumstances, and How? *Terrorism and Political Violence*, 31 (6), 1149–1167.

Gill, P., Corner, E., Conway, M., Thornton, A., Bloom, M., and Horgan, J., 2017. Terrorist Use of the Internet by the Numbers. *Criminology & Public Policy*, 16 (1), 99–117.

Gillespie, T., 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.

Glazzard, A., 2017. Losing the Plot: Narrative, Counter-Narrative and Violent Extremism. *Terrorism and Counter-Terrorism Studies*.

Gorwa, R., 2019. The platform governance triangle: conceptualising the informal regulation of online content. *Internet Policy Review*, 8 (2).

Griffith-Dickson, G., Dickson, A., and Robert, I., 2014. Counter-extremism and De-radicalisation in the UK: a Contemporary Overview. *Journal for Deradicalization*, 0 (1), 26–37.

Guess, A., Nagler, J., and Tucker, J., 2019. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, 5 (1), eaau4586.

Hall, M., M. Logan, G. Ligon, and D. Derrick. 2020. Do Machines Replicate Humans? Towards a Unified Understanding of Radicalizing Content on the Open Social Web. Policy & Internet 12(1): 109–138

Harris-Hogan, S., Barrelle, K., and Zammit, A., 2016. What is countering violent extremism? Exploring CVE policy and practice in Australia. *Behavioral Sciences of Terrorism and Political Aggression*, 8 (1), 6–24.

Heft, A., E. Mayerhöffer, S. Reinhardt, and C. Knüpfer. 2020. Beyond Breitbart: Comparing Right-Wing Digital News Infrastructures in Six Western Democracies. *Policy & Internet* 12(1): 20–45.

Helmus, T.C., 2018. *Russian social media influence: understanding Russian propaganda in Eastern Europe*. Santa Monica, Calif: RAND Corporation.

Holt, T.J., Freilich, J.D., and Chermak, S.M., 2017. Internet-Based Radicalization as Enculturation to Violent Deviant Subcultures. *Deviant Behavior*, 38 (8), 855–869.
Hoskins, A., Awan, A., and O'Loughlin, B., 2011. *Radicalisation and Media: Connectivity and Terrorism in the New Media Ecology*. Routledge.
Hughes, B., 2019. Thriving from Exile: Toward a Materialist Analysis of the Alt-Right. *boundary 2*, 4 (2).
Ingram, H., 2016. *Deciphering the Siren Call of Militant Islamist Propaganda: Meaning, Credibility & Behavioural Change*. The Hague: International Centre for Counter-Terrorism.
Klein, O. and Muis, J., 2019. Online discontent: comparing Western European far-right groups on Facebook. *European Societies*, 21 (4), 540–562.
Lee, B. 2020. Countering Violent Extremism Online: The Experiences of Informal Counter Messaging Actors. *Policy & Internet* 12(1): 66–87.
Lewis, R., 2018. *Alternative Influence: Broadcasting the Reactionary Right on YouTube*. New York: Data & Society Research Institute.
Lewis, R., 2020. "This Is What the News Won't Show You": YouTube Creators and the Reactionary Politics of Micro-celebrity. *Television & New Media*, 21 (2), 201–217.
Marwick, A., 2018. Why Do People Share Fake news? A Sociotechnical Model of Media Effects. *Georgetown Law Technology Review*, 474.
Marwick, A. and Lewis, R., 2017. *Media Manipulation and Disinformation Online*. New York: Data & Society.
Munn, L., 2019. Alt-right pipeline: Individual journeys to extremism online. *First Monday*, 24 (6).
Neumann, P.R., 2013. Options and Strategies for Countering Online Radicalization in the United States. *Studies in Conflict & Terrorism*.
Noble, S.U., 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.
Noelle-Neumann, E., 1991. The Theory of Public Opinion: The Concept of the Spiral of Silence. *Annals of the International Communication Association*, 14 (1), 256–287.
Pearson, E., 2017. Online as the New Frontline: Affect, Gender, and ISIS-Take-Down on Social Media. *Studies in Conflict & Terrorism*, 0 (0), 1–25.
Prucha, N., 2016. IS and the Jihadist Information Highway – Projecting Influence and Religious Identity via Telegram. *Perspectives on Terrorism*, 10 (6).
Reeve, R., 2020. Human Assessment and Crowdsourced Flagging. *In*: B. Ganesh and J. Bright, eds. *Extreme Digital Speech: Contexts, Responses, and Solutions*. Dublin: VOX-Pol.
Richards, I., 2019. A Philosophical and Historical Analysis of "Generation Identity": Fascism, Online Media, and the European New Right. *Terrorism and Political Violence*, 0 (0), 1–20.
Roberts, S.T., 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.
Rudinac, S., Gornishka, I., and Worring, M., 2017. Multimodal Classification of Violent Online Political Extremism Content with Graph Convolutional Networks. *In*: *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*. New York, NY, USA: ACM, 245–252.
Schmitt, J.B., Rieger, D., Rutkowski, O., and Ernst, J., 2018. Counter-messages as Prevention or Promotion of Extremism?! The Potential Role of YouTubeRecommendation Algorithms. *Journal of Communication*, 68 (4), 780–808.

Scrivens, R., Davies, G., and Frank, R., 2018. Searching for signs of extremism on the web: an introduction to Sentiment-based Identification of Radical Authors. *Behavioral Sciences of Terrorism and Political Aggression*, 10 (1), 39–59.
Scrivens, R. and Perry, B., 2017. Resisting the Right: Countering Right-Wing Extremism in Canada. *Canadian Journal of Criminology and Criminal Justice*.
Selim, G., 2016. Approaches for Countering Violent Extremism at Home and Abroad. *The ANNALS of the American Academy of Political and Social Science*, 668 (1), 94–101.
Shehabat, A. and Mitew, T., 2018. Black-boxing the Black Flag: Anonymous Sharing Platforms and ISIS Content Distribution Tactics. *Perspectives on Terrorism*, 12 (1).
Szmania, S. and Fincher, P., 2017. Countering Violent Extremism Online and Offline. *Criminology & Public Policy*, 16 (1), 119–125.
Topinka, R.J., 2018. Politically incorrect participatory media: Racist nationalism on r/ImGoingToHellForThis. *New Media & Society*, 20 (5), 2050–2069.
Vieth, K., 2019. Policing 'online radicalization': the framing of Europol's Internet Referral Unit. *In*: B. Wagner, M. Kettemann, and K. Vieth, eds. *Research Handbook on Human Rights and Digital Technology*. Edward Elgar Publishing.
Vosoughi, S., Roy, D., and Aral, S., 2018. The spread of true and false news online. *Science*, 359 (6380), 1146–1151.
Winter, C., 2019. *Daesh Propaganda, Before and After its Collapse*. Riga: NATO Strategic Communications Centre of Excellence.