



## Countering Extremists on Social Media: Challenges for Strategic Communication and Content Moderation

Bharath Ganesh and Jonathan Bright

---

*Extremist exploitation of social media platforms is an important regulatory question for civil society, government, and the private sector. Extremists exploit social media for a range of reasons—from spreading hateful narratives and propaganda to financing, recruitment, and sharing operational information. Policy responses to this question fit under two headings, strategic communication and content moderation. At the center 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, while being conscious of rights to free expression and the appropriateness of restrictions on speech. This special issue on “Countering Extremists on Social Media: Challenges for Strategic Communication and Content Moderation” focuses on one form of strategic communication, countering violent extremism. In this editorial we discuss the background and effectiveness of this approach, and introduce five articles which develop multiple strands of research into responses and solutions to extremist exploitation of social media. We conclude by suggesting an agenda for future research on how multistakeholder initiatives to challenge extremist exploitation of social media are conceived, designed, and implemented, and the challenges these initiatives need to surmount.*

---

**KEY WORDS:** social media, extremism, regulation, online platforms, counter messaging, content moderation, free speech, countering violent extremism, CVE

极端分子对社交媒体平台的滥用是公民社会、政府和私人部门关注的一个重要监管疑问。极端分子出于一系列原因滥用社交媒体：从散播仇恨叙事和宣传，到金融、招聘和分享操作信息。该疑问所获得的政策响应有两个方面：战略传播和内容审核。这两类政策响应的核心是计算如何最大程度地限制受众接收到极端主义叙事，保持极端主义观点的边缘性，同时考虑自由言论权和言论限制的适当性。本期特刊主题为《打击社交媒体极端主义：战略传播与内容审核面临的挑战》，聚焦于战略传播的一种形式，即反暴力极端主义(CVE)。本期社论中，我们探讨了CVE这种措施的背景和有效性，同时介绍五篇文章，后者就有关极端分子滥用社交媒体的响应与解决措施提出了多个研究。我们的结论为有关如何感知、设计和执行“拒绝极端分子滥用社交媒体”的多个利益攸关方倡议的未来研究提出一项议程，并提出了这些倡议需克服的挑战。

---

**关键词：** 社交媒体，极端主义，监管，网络平台，信息反击，内容审核，自由言论，反暴力极端主义，CVE

*La explotación extremista de las plataformas de redes sociales es una cuestión regulatoria importante para la sociedad civil, el gobierno y el sector privado. Los extremistas explotan las redes sociales por*

*una variedad de razones, desde difundir narrativas de odio y propaganda hasta financiar, reclutar y compartir información operativa. Las respuestas de política a esta pregunta encajan en dos encabezados: comunicación estratégica y moderación de contenido. En el centro de ambas respuestas políticas se encuentra un cálculo sobre la mejor manera de limitar la exposición de la audiencia a las narrativas extremistas y mantener la marginalidad de los puntos de vista extremistas, mientras se tiene conciencia de los derechos a la libertad de expresión y la idoneidad de las restricciones al discurso. Este número especial sobre "Contrarrestar a los extremistas en las redes sociales: desafíos para la comunicación estratégica y la moderación de contenido" se centra en una forma de comunicación estratégica, contrarrestar el extremismo violento (CVE). En este editorial discutimos los antecedentes y la efectividad de este enfoque e introducimos cinco artículos que desarrollan múltiples líneas de investigación sobre respuestas y soluciones a la explotación extremista de las redes sociales. Concluimos sugiriendo una agenda para futuras investigaciones sobre cómo se conciben, diseñan e implementan las iniciativas de múltiples partes interesadas para desafiar la explotación extremista de las redes sociales, y los desafíos que estas iniciativas deben superar.*

---

**PALABRAS CLAVE:** Redes sociales, extremismo, regulación, plataformas en línea, contra mensajes, moderación de contenido, libertad de expresión, contrarrestar el extremismo violento, CVE

## Introduction

Extremist exploitation of social media platforms is an important regulatory question for civil society, government, and the private sector (Crosset & 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, Macdonald, Jarvis, & Chen, 2016; Braddock & Horgan, 2016; Davies, Neudecker, Ouellet, Bouchard, & Ducol, 2016; Ganesh & Bright, 2020; Helmus, 2018; Szmania & Fincher, 2017). 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 & Pearson, 2018), significant attention is being paid to counter-messaging and other strategic communication techniques as potential responses (Bertram, 2016; Beutel et al., 2016; Braddock & Horgan, 2016; Briggs & Feve, 2013; Brown & Marway, 2018; Cherney, 2016; Eerten & van Doosje, 2019). How best to respond to extremism on social media often centers 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 programs 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 although 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 center 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 (Aly, Balbi, & Jacques, 2015; Briggs & Feve, 2013; Brown & Marway, 2018; Dalgaard-Nielsen, 2016; Gielen, 2019; Griffith-Dickson, Dickson, & Robert, 2014; Scrivens & Perry, 2017). 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 multistakeholder 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**

CVE refers to a field of “soft power” mechanisms that try to counter extremists, and which 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 from radicalizing influences (Bjola & Pamment, 2019, p. 7; Selim, 2016, p. 95). Alongside working with local communities and supporting individuals, strategic communications is one of the key functions of CVE.<sup>1</sup> Many CVE programs 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 United Kingdom (described below). Broadly, CVE initiatives incorporate contributions from civil society, governments, think tanks and non-profits, and the private sector.

CVE activities can be conceptualized as primary, secondary, and tertiary. Primary CVE seeks to reduce the likelihood of radicalization across a population, secondary CVE focuses on those vulnerable to radicalization, and tertiary CVE focuses on those already radicalized (Gielen, 2019, p. 1157; Harris-Hogan, Barrelle, & Zammit, 2016). While secondary and tertiary CVE often involve state-run exit and deradicalization programs, 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,<sup>2</sup> is working with civil society for counter-messaging and anti-hate work,<sup>3</sup> and moderates content and suspends users where necessary.<sup>4</sup> 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 U.K. Home Office's Research, Information and Communications Unit (RICU) contracting Breakthrough Media,<sup>5</sup> 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,<sup>6</sup> which has continued in the United Kingdom under the "Building a Stronger Britain Together" program run by the Home Office.<sup>7</sup> While it remains to be seen what effect this investment has had on preventing extremism and disrupting circuits of radicalization, 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 (Archetti, 2019; Awan, Miskimmon, & O'Loughlin, 2019; Glazzard, 2017). When used in conjunction with automated recommendation systems, they may even risk counter-productive effects (Bright, Marchal, Ganesh, & Rudinac, 2020; Schmitt, Rieger, Rutkowski, & Ernst, 2018).

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 (2020) argues that overt participation from a government may compromise the credibility of CVE activity (see also Belanger & Szmania, 2018; Ingram, 2016; Neumann, 2013). 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 Benjamin Lee's work on informal counter-narratives and Irfan Chaudhry and Anatoliy Gruzd's work on comment section racism on Facebook news pages. Where Bilazarian develops policy recommendations that can better guide online-oriented CVE, Lee (2020) and Chaudhry and Gruzd (2020) 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?". 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 behaviors." 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 satirizing, criticizing, 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 to 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 (2020) 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" (Noelle-Neumann, 1991). Although 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 viewpoints 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 labor 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 Maura Conway and colleagues find that Twitter takedown of pro-IS accounts “severely affected IS’s ability to develop and maintain robust and influential communities on Twitter” (Berger & 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 decreased (Chandrasekharan et al., 2017). However, 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-IS 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, 2018).

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; Reeve, 2020; Vieth, 2019). Further, social media companies have developed their own relationship with specific civil society organizations 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 organizations, think tanks, and civil society (Gorwa, 2019). More recently, there have 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 (Borisyyuk, Gordo, & Sivakumar, 2018; Burnap & Williams, 2016; Djuric et al., 2015; Rudinac, Gornishka, & Worrington, 2017; Scrivens, Davies, & Frank, 2018). These 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, Logan, Ligon, and Derrick (2020) 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. (2020) provide valuable advancements in research design and methodology that can be applied for future study; probing the possibilities and limits of technical systems in primary CVE, and 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 which provides an important resource for extremists, this ecosystem presents significant challenges in ensuring the viability of primary CVE activities. Research on disinformation and polarization on social media has highlighted the important roles played by users in spreading “fake news” (Vosoughi, Roy, & Aral, 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, Nagler, & Tucker, 2019), and the disproportionate role of radical right media in the spread of disinformation (Bennett & Livingston, 2018). Recent work has stressed the central role that alternative media—such as Breitbart News and the use of social media by 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, Faris, & Roberts, 2018; Bennett & 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 antidemocratic, repressive, and denigrating to out-groups while continuing to challenge hegemonic discourse in mainstream media (Heft, Mayerhöffer, Reinhardt, & Knüpfer, 2020). 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 hyperpartisan alternative media is central to primary CVE; consequently, an engagement (offered by Heft et al., 2020, 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 legitimizing 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 (Hughes, 2019; Lewis, 2018; Marwick, 2018; Munn, 2019).

While public attention has shifted to ISIS media in recent years, jihadists have long developed news outlets, web forums, and networks to share training and technical manuals (Awan, 2007, pp. 76–77; Archetti, 2012; Conway, McInerney, & Ducol, 2012; Hoskins, Awan, & O'Loughlin, 2011). 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 United Kingdom (Awan, 2007, p. 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 (Al-Rawi, 2018; Baele, Boyd, & Coan, 2019; Fisher, 2015; Fisher, Prucha, & Winterbotham, 2019; Shehabat & Mitew, 2018; 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.<sup>8</sup>

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 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.<sup>9</sup> 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, Lewis, & Friedberg, 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 (Hall et al., 2020). While research on far-right exploitation of social media is increasing, much of this work focuses on

representation, narratives, ideology, and discourse (Deem, 2019; Froio & Ganesh, 2019; Klein & Muis, 2019; Richards, 2019; Topinka, 2018), as well as disinformation spreading from right-wing digital news to social media platforms and mainstream media (Benkler et al., 2018; Bennett & Livingston, 2018; Marwick & Lewis, 2017). Developing the context of such activity in comparative perspective, Heft et al. (2020) in this special issue provide a thorough mapping of hyperpartisan 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; finding various structures, styles, and supply and demand markets.

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. (2020) 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. 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 hyperpartisan from regular news (Heft et al., 2020). 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. (2020) explore cannot be uniformly or uncontroversially referred to as extreme right, nor are they directly implicated in far-right terrorism, they do demonstrate a number of significant trends that are 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 nonviolent extremist subcultures can thrive (see Holt, Freilich, & Chermak, 2017). While there is little research to prove that these nonviolent 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 media can undermine efforts at primary CVE and must be understood as actors that present a challenge to both formal and informal counter-messaging. This last article in the 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 right-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 editorial, 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 these 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 decidedly 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.

**Bharath Ganesh**, Centre for Media and Journalism Studies, University of Groningen, Groningen, The Netherlands and Oxford Internet Institute, University of Oxford, Oxford, UK [b.ganesh@rug.nl].

**Jonathan Bright**, Oxford Internet Institute, University of Oxford, Oxford, UK.

### Notes

This work was financed by the VOX-Pol Network, which is funded by the EU 7th Framework Programme (grant number 312827).

1. The U.S.-led Global Engagement Centre, the U.K. 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).
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>
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](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/836780/building-stronger-britain-together-2019-horr112.pdf)
8. <https://gifct.org/press/global-internet-forum-counter-terrorism-update-our-progress-two-years/>
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](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)

## References

- Al-Rawi, A. 2018. "Video Games, Terrorism, and ISIS's Jihad 3.0." *Terrorism and Political Violence* 30 (4): 740–60.
- Aly, A., A.-M. Balbi, and C. Jacques. 2015. "Rethinking Countering Violent Extremism: Implementing the Role of Civil Society." *Journal of Policing, Intelligence and Counter Terrorism* 10 (1): 3–13.
- Aly, A., S. Macdonald, L. Jarvis, and T. Chen. 2016. *Violent Extremism Online: New Perspectives on Terrorism and the Internet*. New York, NY: Routledge.
- Archetti, C. 2012. *Understanding Terrorism in the Age of Global Media: A Communication Approach*. New York, NY: Palgrave Macmillan.
- Archetti, C. 2019. "The Unbearable Thinness of Strategic Communication." In *Countering Online Propaganda and Extremism: The Dark Side of Digital Diplomacy*, eds. C. Bjola, and J. Pamment. London: Routledge, 81–96.
- Awan, A., A. Miskimmon, and B. O'Loughlin. 2019. "The Battle for the Battle of the Narratives: Side-Stepping The Double Fetish of Digital and CVE." In *Countering Online Propaganda and Extremism: The Dark Side of Digital Diplomacy*, eds. C. Bjola, and J. Pamment. London: Routledge, 156–71.
- Awan, A.N. 2007. "Radicalization on the Internet?" *The RUSI Journal* 152 (3): 76–81.
- Baele, S.J., K. Boyd, and T. Coan. 2019. *ISIS Propaganda: A Full-Spectrum Extremist Message*. Oxford: Oxford University Press.
- Belanger, P., and S. Szmania. 2018. "The Paradox of Source Credibility in Canadian and U.S. Domestic Counterterrorism Communications." *International Journal of Communication* 12: 22.
- Benkler, Y., R. Faris, and H. Roberts. 2018. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford: Oxford University Press.
- Bennett, W.L., and S. Livingston. 2018. "The Disinformation Order: Disruptive Communication and the Decline of Democratic Institutions." *European Journal of Communication* 33 (2): 122–39.
- Berger, J.M., and H. Perez. 2016. *The Islamic State's Diminishing Returns on Twitter: How Suspensions are Limiting the Social Networks of English-speaking ISIS Supporters*. Washington, DC: 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* 7: 225–52.
- Beutel, A., S. Weine, A. Saeed, A. Mihajlovic, A. Stone, J. Behrs, and S. Shanfield. 2016. "Guiding Principles for Countering and Displacing Extremist Narratives." *Contemporary Voices: St Andrews Journal of International Relations* 7 (3): 35.
- Bilazarian, T. 2020. "Countering Violent Extremist Narratives Online: Lessons From Offline Countering Violent Extremism." *Policy & Internet* 12 (1): 46–65.
- Bjola, C., and J. Pamment. 2019. "Introduction: the 'Dark Side' of Digital Diplomacy." In *Countering Online Propaganda and Extremism: The Dark Side of Digital Diplomacy*, eds. C. Bjola, and J. Pamment. London: Routledge, 1–10.
- Borisyuk, F., A. Gordo, and V. Sivakumar, 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, UK: ACM Press, 71–79.
- Braddock, K., and J. Horgan. 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 S. Feve. 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.
- Bright, J., Marchal, N., Ganesh, B., and Rudinac, S. 2020. "Echo Chambers Exist! (But They're Full of Opposing Views)." <https://arxiv.org/abs/2001.11461>

- Brocato, A. 2015. "Tackling Terrorists' Use of the Internet: Propaganda Dispersion & the Threat of Radicalization." In *Terrorist Use of Cyberspace and Cyber Terrorism: New Challenges and Responses*, ed. M.N. Ogun. Amsterdam: IOS Press, 129–48.
- Brown, K., and H. Marway, 2018. *Preventing Radicalisation to Terrorism and Violent Extremism: Delivering counter- or alternative narratives*. Radicalisation Awareness Network.
- Brown, K., and E. Pearson. 2018. "Social Media, The Online Environment and Terrorism." In *Routledge Handbook of Terrorism and Counterterrorism*, ed. A. Silke. London: Routledge.
- Burnap, P., and M.L. Williams. 2016. "Us and Them: Identifying Cyber Hate on Twitter Across Multiple Protected Characteristics." *EPJ Data Science* 5 (1): 11.
- Chandrasekharan, E., U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert. 2017. "You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech." *Proceedings of the ACM on Human-Computer Interaction* 1: 1–22.
- 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. Gruz. 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". *SSRN Scholarly Paper No. ID 2941880*. Rochester, NY: Social Science Research Network.
- Conway, M., M. Khawaja, S. Lakhani, J. Reffin, A. Robertson, and D. Weir. 2019. "Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts." *Studies in Conflict & Terrorism* 42 (1–2): 141–60.
- Conway, M., L. McInerney, and B. Duco. 2012. "Uncovering the French-Speaking Jihadisphere: An Exploratory Analysis." *Media, War & Conflict* 5 (1): 51–70.
- Coyer, K. 2020. "Informal Counter-Narratives." In *Extreme Digital Speech: Contexts, Responses, and Solutions*, eds. B. Ganesh, and J. Bright. Dublin: VOX-Pol.
- Crosset, V., and B. Dupont. 2018. "Internet et propagande jihadiste: la régulation polycentrique du cyberspace." *Critique Internationale* 78 (1): 107–25.
- Dalgaard-Nielsen, A. 2016. "Countering Violent Extremism With Governance Networks." *Perspectives on Terrorism* 10 (6): 135–39.
- Davies, G., C. Neudecker, M. Ouellet, M. Bouchard, and B. Duco. 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: 20.
- Djuric, N., J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, 2015. "Hate Speech Detection with Comment Embeddings." In: *Proceedings of the 24th International Conference on World Wide Web*. New York, NY: ACM, 29–30.
- Donovan, J., B. Lewis, and B. Friedberg. 2018. "Parallel Ports: Sociotechnical Change From the Alt-Right to Alt-Tech." In *Post-Digital Cultures of the Far Right: Online Actions and Offline Consequences in Europe and the US*, eds. M. Fielitz, and N. Thurston. Bielefeld: transcript Verlag.
- Eerten, J.-J., and B. van Doosje. 2019. *Challenging Extremist Views on Social Media: Developing a Counter-Messaging Response*. Abingdon: Routledge.
- Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Publishing Group.
- Fisher, A. 2015. "How Jihadist Networks Maintain a Persistent Online Presence." *Perspectives on Terrorism* 9 (3): 3–20.
- Fisher, A., N. Prucha, and E. 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 B. Ganesh. 2019. "The Transnationalisation of Far Right Discourse on Twitter." *European Societies* 21 (4): 513–39.
- Gallacher, J. 2020. "Automated Detection of Terrorist and Extremist Content." In *Extreme Digital Speech: Contexts, Responses, and Solutions*, eds. B. Ganesh, and J. Bright. Dublin: VOX-Pol.
- Ganesh, B., and J. Bright. 2020. "Introduction." In *Extreme Digital Speech: Contexts, Responses and Solutions*, eds. B. Ganesh, and J. Bright. Dublin: VOX-Pol Network of Excellence, 10–14.
- 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): 1149–67.
- Gill, P., E. Corner, M. Conway, A. Thornton, M. Bloom, and J. Horgan. 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): 1–22.
- Griffith-Dickson, G., A. Dickson, and I. Robert. 2014. "Counter-Extremism and De-Radicalisation in the UK: A Contemporary Overview." *Journal for Deradicalization* 1: 26–37.
- Guess, A., J. Nagler, and J. Tucker. 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., K. Barrelle, and A. Zammit. 2016. "What is Countering Violent Extremism? Exploring CVE Policy and Practice in Australia." *Behavioral Sciences of Terrorism and Political Aggression* 8 (1): 6–24.
- Helmus, T.C. 2018. *Russian Social Media Influence: Understanding Russian Propaganda in Eastern Europe*. Santa Monica, CA: RAND Corporation.
- 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.
- Holt, T.J., J.D. Freilich, and S.M. Chermak. 2017. "Internet-Based Radicalization as Enculturation to Violent Deviant Subcultures." *Deviant Behavior* 38 (8): 855–69.
- Hoskins, A., A. Awan, and B. O'Loughlin. 2011. *Radicalisation and Media: Connectivity and Terrorism in the New Media Ecology*. London: 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 J. Muis. 2019. "Online Discontent: Comparing Western European Far-Right Groups on Facebook." *European Societies* 21 (4): 540–62.
- 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, NY: 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): 201–17.
- Marwick, A. 2018. "Why Do People Share Fake News? A Sociotechnical Model of Media Effects." *Georgetown Law Technology Review* 2: 474–512.
- Marwick, A., and R. Lewis. 2017. *Media Manipulation and Disinformation Online*. New York, NY: 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* 36: 431–59.
- 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): 256–87.
- Pearson, E. 2018. "Online as the New Frontline: Affect, Gender, and ISIS-Take-Down on Social Media." *Studies in Conflict & Terrorism* 41 (11): 850–74.
- Prucha, N. 2016. "IS and the Jihadist Information Highway—Projecting Influence and Religious Identity via Telegram." *Perspectives on Terrorism* 10 (6): 54.
- Reeve, R. 2020. "Human Assessment and Crowdsourced Flagging." In *Extreme Digital Speech: Contexts, Responses, and Solutions*, eds. B. Ganesh, and J. Bright. 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*, 1–20. <https://www.tandfonline.com/doi/abs/10.1080/09546553.2019.1662403?journalCode=ftpv20>
- Roberts, S.T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. London: 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: ACM, 245–52.
- Schmitt, J.B., D. Rieger, O. Rutkowski, and J. Ernst. 2018. "Counter-Messages as Prevention or Promotion of Extremism?! The Potential Role of YouTube Recommendation Algorithms." *Journal of Communication* 68 (4): 780–808.
- Scrivens, R., G. Davies, and R. Frank. 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 B. Perry. 2017. "Resisting the Right: Countering Right-Wing Extremism in Canada." *Canadian Journal of Criminology and Criminal Justice* 59: 534–58.
- 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 T. Mitew. 2018. "Black-Boxing the Black Flag: Anonymous Sharing Platforms and ISIS Content Distribution Tactics." *Perspectives on Terrorism* 12 (1): 81–99.
- Szmania, S., and P. Fincher. 2017. "Countering Violent Extremism Online and Offline." *Criminology & Public Policy* 16 (1): 119–25.
- Topinka, R.J. 2018. "Politically Incorrect Participatory Media: Racist Nationalism on r/ImGoingToHellForThis." *New Media & Society* 20 (5): 2050–69.
- Vieth, K. 2019. "Policing 'Online Radicalization': The Framing of Europol's Internet Referral Unit." In *Research Handbook on Human Rights and Digital Technology*, eds. B. Wagner, M. Kettemann, and K. Vieth. Cheltenham, UK: Edward Elgar Publishing.
- Vosoughi, S., D. Roy, and S. Aral. 2018. "The Spread of True and False News Online." *Science* 359 (6380): 1146–51.
- Winter, C. 2019. *Daesh Propaganda, Before and After Its Collapse*. Riga: NATO Strategic Communications Centre of Excellence.