EXTREME DIGITAL SPEECH

CONTEXTS, RESPONSES AND SOLUTIONS

Edited by Bharath Ganesh and Jonathan Bright

EXTREME DIGITAL SPEECH

CONTEXTS, RESPONSES AND SOLUTIONS

Acknowledgements
This report is an output from a workshop organised jointly by VOX-Pol and the Oxford Internet Institute in May 2018. The workshop was titled 'Combating Online Extremism: State, Private Sector and Civil Society Responses' and took place at St Antony's College, Oxford. The report's contributors and editors thank VOX-Pol for funding and helping organise the workshop. We are particularly grateful for support from Shahed Warreth, Sadhbh Byrne Ryan, Lisa McInerney and Maura Conway, and other VOX-Pol team members at Dublin City University.

Both Dr Joel Busher (Coventry) and Dr Benjamin Lee (Lancaster) provided detailed reviews of all of the articles at their initial stages, encouraging valuable revisions to all of the pieces in this collection. We are very grateful for their participation in this project. In addition to their effort, this collection benefitted greatly from the efforts of two anonymous peer reviewers whose detailed comments have helped to improve the contributions from all the authors.

ISBN: 978-1-873769-96-6

Designed and typeset by Soapbox, www.soapbox.co.uk

TABLE OF CONTENTS

PART II. COUNTERING EXTREME DIGITAL SPEECH: RESPONSES AND SOLUTIONS

CONTRIBUTORS

Mubaraz Ahmed is a senior analyst in the Co-Existence team at the Tony Blair Institute for Global Change, where he focuses on Islamist extremism. He also works closely with the Institute's projects and education teams, providing insight and expertise to better inform programmatic work. Mubaraz leads the Institute's work into online extremism, having co-authored a groundbreaking report into the accessibility of Islamist extremist content in Google search results, *A War of Keywords*. He regularly engages with governments and technology companies to support the development of sustainable policy and practical solutions to the challenges of online extremism.

Laurence Bindner delivers expertise on the organisation of jihadist online activity and the spread of content, providing knowledge on the evolution of jihadi narratives and rhetoric. Former Director of Development of the Center for the Analysis of Terrorism (CAT) in Paris and member of the United Nations Counter-Terrorism Committee and its Executive Directorate (UN CTED) Global Research Network, her work also covers analysis of terrorism financing.

Jonathan Bright is a Senior Research Fellow at the Oxford Internet Institute who specialises in computational approaches to understanding online political behaviour and communication. His research interests include the study of digital campaign communication, echo chambers, online radical behaviour and online harms. His recent work has been published in leading journals in the field such as *Journal of Communication, Communication Research and Journal of Computer-Mediated Communication*. He is also an editor of the journal *Policy & Internet*.

Joel Busher is an Associate Professor in the Centre for Trust, Peace and Social Relations at Coventry University. He has published widely on anti-minority activism, far-right extremism, the social ecology of political violence and the implementation of counter-terrorism policy. Joel's 2016 monograph, *The Making of Anti-Muslim Protest: Grassroots Activism in the English Defence League*, was awarded the British Sociological Association's Philip Abrams Memorial Prize and he is currently leading numerous projects on anti-minority activism and right-wing extremism. His recent work can be found in leading journals including *Social Movement Studies, Critical Studies on Terrorism and Sociology*.

Kate Coyer is a Research Affiliate with the Centre for Media, Data and Society (CMDS) in the School of Public Policy at Central European University (CEU), and an Affiliate of the Berkman Klein Center for Internet and Society at Harvard University. She leads CMDS's work related to technology and civil society, and she was previously executive director of the Centre. Kate completed her PhD in Media and Communications at Goldsmiths College and a postdoctoral fellowship at the University of Pennsylvania. Recently, she directed CEU work on the Virtual Centre of Excellence for Research in Violent Online Political Extremism (VOX-Pol), which explores the role of social media platforms in responding to extremism online and the complexities of the relationship between technology and free expression that lie at the intersection of global security and human rights.

Valentine Crosset is a PhD candidate in Criminology at Université de Montréal. She is interested in online manifestations of extremism, and the technology and regulation that frame responses to online hate speech and extremism. In her dissertation, Valentine explores the visibility of extremist groups on the Internet. Furthermore, she examines the latest techniques, methods and innovative tools used by violent communities such as Islamist groups to increase their online visibility and circumvent moderation. Her work has been published in peer-reviewed journals such as *New Media & Society* and *Critique internationale*.

Huw Davies is a Researcher at the Oxford Internet Institute, where he studies the co-constitutive relationships between digital technologies and social inequalities. His PhD, completed at the University of Southampton, explored the sociology of young people's digital literacies. Huw has contributed to numerous research projects at the Institute, focusing on social inequalities in research areas including crowdwork, child safety online, Internet access and media literacy. He is also a co-convenor of the British Sociological Association's Digital Sociology Study Group. His work has been published in journals including *Information, Communication & Society, New Media & Society and Sociological Research Online*.

John Gallacher is a DPhil student within the University of Oxford's Cyber Security Centre for Doctoral Training. His research focuses on investigating the causes of online polarisation, ranging from aggressive intergroup contact, the spread of extremist material and hostile interference from foreign states. John's work combines analytic methods drawn from computer science (machine learning, natural language processing and network science) with insights from experimental psychology and open-source information from social media in order to measure how groups interact online, and how this relates to real-world events.

Bharath Ganesh is an Assistant Professor in Media Studies at the University of Groningen and a Visiting Research Associate at the Oxford Internet Institute. His research focuses on hate, intolerance and prejudice in Europe and North America, racism and bigotry in digital propaganda, and regulatory responses to online extremism. Bharath was a postdoctoral researcher at the Oxford Internet Institute, where he contributed to projects on computational propaganda, the VOX-Pol Network of Excellence, and projects on data science use in local government. He holds a PhD in Geography from University College London. His recent publications have appeared in *Journal of International Affairs, Foreign Policy and European Societies*.

Raphael Gluck has a background in web development, social media marketing and IT consultancy. Over the past several years he has focused on the spread of terrorist propaganda online, charting jihadist group digital strategies, including app development and social media proliferation on the surface web as well as the deep web and the dark web.

Benjamin Lee is a Senior Research Associate at the Centre for Research and Evidence on Security Threats (CREST) at the University of Lancaster. At the University of Leicester, Benjamin contributed to projects on far-right communication. And at the University of Northampton, he helped collect oral histories and archival materials on anti-fascist campaigning in the UK. His recent work has focused on online counter-jihad networks, informal counter-messaging and countering violent extremism. His recent publications can be found in leading journals including *Policy & Internet, Studies in Conflict & Terrorism and International Journal of Press/Politics*.

Matti Pohjonen is a Lecturer in Global Digital Media at the School of Oriental and African Studies (SOAS), University of London. For the past ten years, he has developed critical-comparative research approaches for understanding digital cultures globally. This has included work on international news and blogging in India, mobile technology in East Africa, and comparative research on online extremism and hate speech in Ethiopia and Europe, as well as exploring new methods in Big Data analysis for digital media research. He received his PhD from SOAS, where he also worked as a Senior Teaching Fellow and an AHRC-funded Post-Doctorate Research Fellow. He has also held research positions at the University of Oxford, University of Bremen, the University of Helsinki and the VOX-Pol Network of Excellence. His recent publications include contributions to *International Journal of Communication* and the VOX-Pol report *Horizons of Hate*.

Zoey Reeve has a background in psychology, terrorism studies and political science. Her research focuses on the social-evolutionary psychology of radicalisation and terrorism in both online and offline spheres. She has a particular interest in individual differences in the radicalisation process. Zoey received her PhD in Political Science from the University of Edinburgh and is a VOX-Pol Research Fellow. Her recent publications have appeared in the journal *Terrorism and Political Violence* and include studies that use innovative experimental paradigms to generate new data on the radicalisation process.

INTRODUCTION

Bharath Ganesh and Jonathan Bright

Extreme digital speech (EDS) is an emerging challenge that requires co-ordination between governments, civil society and the private sector. In this report, a range of experts on countering extremism consider the challenges that EDS presents to these stakeholders, the impact that EDS has and the responses taken by these actors to counter it. By focusing on EDS, our consideration of the topic is limited to the forms of extreme speech that take place online, often on social media platforms and multimedia messaging applications such as WhatsApp and Telegram. Furthermore, by focusing on EDS rather than explicitly violent forms of extreme speech online, we (as Matti Pohjonen writes in this report) 'depart' from a focus on violence and incorporate a broader range of issues such as hateful and dehumanising speech and the complex cultures and politics that have formed around EDS. This focus brings into view a much broader range of factors that help assemble extremist networks online. This perspective is necessary, given the role that hate speech plays in extreme right-wing networks and the complexity of Daesh propaganda which uses videos to synthesise utopic images of life in the so-called 'Khilafa'. Following JM Berger's recent book, Extremism (2018), we can think of EDS as a core practice that produces an archive of online extremist resources that reinforce the sense of in-group belonging across a network of geographically dispersed users, whether this be the networks of jihadists connected on Telegram, or right-wing extremists that use trolling tactics to hack mainstream opinion on Twitter.

All the same, while it is well-known that EDS is prolific online, there is little understanding of what kind of impact participation in these networks actually has on the likelihood of an individual's engagement in political violence. Moreover, very little is known about what methods are being used to challenge EDS and what solutions are

best suited for this problem. This report seeks to provide policymakers, researchers and practitioners with an overview of the context of EDS, its impact, and the responses and solutions being mobilised to counter it. In order to do this, this report assembles a set of ten brief essays intended to serve as a starting point for further exploration of a range of topics related to the management of EDS across government, civil society and the private sector.

The report begins with a contribution by Matti Pohjonen that argues the complexity of EDS requires scholars and practitioners to look beyond narrow definitions of 'extremism' in the boundaries set by legal categories such as 'violence' and 'hate speech'. His piece encourages us to think about the cultures that underwrite the 'vitriol' that straddles the line between acceptable and unacceptable speech, often using irony and humour common to Internet culture. Taking these complexities into account guides our consideration of the 'complex politics' that not only form around EDS, but also the difficulties that government, civil society and private sector actors face in challenging it.

The report is then split into two parts. Part 1, 'Extreme Digital Speech: Contexts and Impact', presents three contributions exploring the context of EDS amongst jihadists and the extreme right in Europe and North America. The first two essays in Part 1 take a high-level view on broad trends for both groups. We focus on jihadist and right-wing extremism because they have motivated numerous attacks across the world, often rely on digital communications, and are a central focus in both global and national counter-extremism agendas. In their essay on jihadist extremism, Laurence Bindner and Raphael Gluck illuminate how jihadist communications online have evolved using a variety of platforms, focusing today on Telegram as a key platform for disseminating jihadist messaging, resources and propaganda. Next, drawing on the complexities of EDS and the difficulty of defining extremism in the context of the extreme right, Bharath Ganesh explores three configurations of the extreme right in the history of interactive online communication, exploring the role of and opportunities provided by webforums, political blogs and, more recently, social media. In the third essay in this section, Mubaraz

Ahmed takes a broader view on the impact of EDS on violence. In reviewing literature on the topic and exploring four cases of violent extremists, Ahmed finds that the effects of extremist content accessible online cannot be predetermined. Because extremist content is used by individuals in a myriad of ways that differ on a case-by-case basis, Ahmed recommends a focus on how EDS is used by violent extremists and why it plays a role in motivating violence. While there is no proven link that coming across EDS online leads to participation in violence, Ahmed suggests that many of the responses – such as account suspensions, takedowns, as well as counter-narratives – are, at times, based on the flawed assumption that EDS and violence are causally linked.

Following from the discussion of context and impact, Part 2, 'Countering Extreme Digital Speech: Responses and Solutions', covers a range of topics focusing on how governments, civil society and the private sector have responded to EDS. This section aims to critically evaluate these approaches. One of the most prominent approaches to countering EDS is automated detection and content removal. In his explanation of the machine learning techniques used in this practice, John Gallacher provides an accessible and detailed explanation of how automated content detection works. He also takes a critical look at the potential for false positives at the massive scale that these technologies may be applied. Zoey Reeve explores human-mediated content assessment and crowdsourced reporting in the area of extremism, exploring how Internet Referral Units (IRUs), particularly in the United Kingdom, play a crucial role in addressing EDS. Nevertheless, Reeve finds that the challenges raised by the subjective nature of moderating content online and the lack of transparency by major platforms persist. Building on Gallacher's and Reeve's contributions, Valentine Crosset considers how responses are formulated across a network of government and private sector actors, ranging from account takedown and suspension, and deletion of content, as well as the context of increasingly demanding laws, such as Germany's NetzDG, that require stricter moderation from social media platforms. However, the fluidity of the definitions of 'extremism' and 'hate speech', alongside the norms that

influence how social media platforms differentiate the content that is acceptable from that which is not, remain difficult challenges for this network of actors to surmount. Furthermore, using more negative measures such as removing and blocking content can have unintended consequences, such as migration to other platforms, despite the efficacy it can have in disrupting networks of EDS.

Whereas Crosset's discussion focuses on negative measures, such as removing and blocking content, the next two essays explore counter-narratives. These represent a positive approach that challenges EDS rather than a negative one that seeks to take it down. Much promise – as well as effort by government, civil society and the private sector – has been attributed to counter-narratives. Reviewing recent evaluations of counter-narratives, Bharath Ganesh suggests that the promise placed in counter-narratives appears to be overstated. This is partially because many counter-narrative programmes have not taken into account many of the best practices recommended by academics and researchers in the area. He stresses that it is important to be more cautious of the promise of counter-narratives, and that, while they most certainly cannot be ignored, they should not form a *central* part of a strategy to counter EDS. Taking a slightly different approach, Kate Coyer looks at the promise of *informal* counter-narratives that are not programmes run by civil society or associated with government, and suggests that their potential to use authentic voices to cast a critical light on extremism may be more promising. Nevertheless, both pieces stress that many of the metrics currently used to evaluate counter-narrative programmes tend to provide only a surface-level understanding of their impact, and that much more research is needed to explore how, if at all, counter-narratives might lead to behavioural change.

In the final essay in this report, Huw Davies critically analyses the confidence placed in digital literacy as a way of building resilience to EDS. Drawing on research in digital literacy and placing it in a historical context, Davies suggests that digital literacy, framed as a discourse of 'upskilling' Internet users, is too narrow a formulation to counter extremism. He argues that it does not adequately recognise that many of those responsible for toxifying digital public discourse

are in fact highly literate digital entrepreneurs. Davies's interrogation of the frameworks of digital literacy in the context of skills unfolds a few pathways to think critically about the limits to digital literacy as a response to extremism.

The ten essays compiled in this report explore how actors across government, civil society, and the private sector have set up regulatory systems to manage and counter extreme digital speech. The reflexive approach to the topic developed by each of the authors should give readers an understanding of the shortcomings and opportunities in the different ways that EDS can be challenged. By presenting critiques and reflections on this topic, the report hopes to give researchers and practitioners an overview of the main challenges in addressing EDS and concise reviews of key debates in the field.

EXTREME DIGITAL SPEECH

Matti Pohjonen

Understanding the growing problem of extremist content online is one of the most contentious issues facing contemporary societies around the world. Scholarly research on violent online political extremism has conventionally approached this problem by exploring how the *online* activities of violent political extremist or terrorist groups contribute to political violence *offline*. This has enabled the research to adopt a relatively easy-to-define normative division between what is considered legitimate forms of political expression (protected under freedom of speech) and what should be criminalised (such as calls to political violence). Brown and Cowls (2015, p. 23) note that "there is a reasonable consensus amongst states and international organisations that inciting a terrorist offence, recruiting, and training for terrorism should be criminalised." However, they also warn that "there is less consensus on what constitutes online 'extremist' material that should be policed – especially where it does not directly encourage violence" (2015, p. 29).

Approaching online extremism through a legal-normative framework centred around a discourse of terrorism and political violence raises two problems highlighted in this chapter. First, the assortment of new forms of online activity that has emerged in recent years defies such easy categorisation into speech that is acceptable and speech that is not. In other words, unlike more traditional forms of violent extremist activity online, new movements around anti-immigrant populism, social media hate speech or the so-called alt-right cannot be as easily typecast into binary divisions between legitimate or illegitimate forms of political speech.

And second, given the close historical relationship between debates on violent online political extremism and the discourse of terrorism especially in the West, researchers working on violent online extremism have often presupposed a universalising normative

framework towards their object of study, which is not as easily transferrable to other examples of online extremism across a diversity of global contexts. Conway (2017, p. 16) writes that "widening our present narrow focus on violent online jihadism to encompass a variety of other contemporary violent online extremisms will allow us to engage in much-needed cross-ideological analysis." Consequently, there is a growing need to better understand the multiplicity of situated speech acts and cultures of communication underlying violent online extremism in countries with often radically different socio-political contexts and 'polymedia' environments (Madianou and Miller 2012).

Regarding the concept of 'extreme speech', Udupa and Pohjonen (2019) emphasise the variation in context and cultural resources of approval behind the many forms of online vitriol. They also show how new movements of right-wing populism and anti-immigrant sentiments globally masquerade their violent messages behind a subterfuge of humour, irony, memes and a style of communication more commonly associated with Internet culture rather than with traditional forms of militant extremism. They write that:

> *Debates around violent online political extremism, and especially 'terrorism talk' popular in the public and political imaginations of online extremism, have revolved around notions of risk and processes of radicalization … [H]owever, it is important to problematize the orthodox understanding of extremism premised on a clear-cut normative binary between the liberal center and the extreme periphery and to explore how these political inclusions and exclusions are themselves produced globally across a range of cultural and political registers.*
>
> **2019, p. 3,051**

This report similarly uses the term 'extreme digital speech' to discuss a diverse range of examples of online extremism – and measures adopted to prevent it – from the various perspectives adopted towards this object of analysis. Leaving the definition of online extremism as open-ended as possible, this report seeks to avoid "predetermining the effects of online volatile speech as vilifying, polarizing, or lethal" (Pohjonen and Udupa 2017, p. 1,174), and enables the report's contributions to depart from the dominant discourse of terrorism and securitisation still often associated with debates on violent online political extremism. Instead, it brings into focus a diversity of perspectives relevant to understanding the problem of online extremism around the world and the complex politics that have developed around it in the recent years.

References

Brown, I. and Cowls, J. 2015. 'Check the Web: Assessing the Ethics and Politics of Policing the Internet for Extremist Material'. *VOX-Pol Network of Excellence*.

Conway, M. 2017. 'Determining the Role of the Internet in Violent Extremism and Terrorism: Six Suggestions for Progressing Research'. *Studies in Conflict & Terrorism*, 40 (1), pp. 77–98.

Madianou, M. and Miller, D. 2012. 'Polymedia: Towards a New Theory of Digital Media in Interpersonal Communication'. *International Journal of Cultural Studies*, 16 (2), pp. 169–187.

Pohjonen, M. and Udupa, S. 2017. 'Extreme Speech Online: An Anthropological Critique of Hate Speech Debates'. *International Journal of Communication,* 11, pp. 1,173–1,191.

Udupa, S. and Pohjonen, M. 2019. 'Extreme Speech and Global Digital Cultures'. *International Journal of Communication*, 13, pp. 3,049–3,067.

PART I. EXTREME DIGITAL SPEECH

CONTEXTS AND IMPACT

JIHADIST EXTREMISM

Laurence Bindner and Raphael Gluck

As early as the 1990s, jihadists understood how leveraging the Internet could further their goals. The Internet's reach and speed, the level of interaction it provides, and its low cost have influenced the paradigm of local empowerment and leaderless jihad (Lia 2008). Jihadists' early forays online had a dual purpose: first, sharing propaganda, strategic lines, specific threats and messages; and, second, operational use with the Internet facilitating communication, sharing of training material, fundraising and recruitment, among others. This essay provides an overview of jihadist groups have used the Internet. After initially looking at the genesis of this use of the Internet by al-Qaeda, we will focus on the recent trends on how jihadists build agile and resilient online content and how the response to them is structured, both at the institutional level and in the private sector.

The first jihadist websites surfaced around 2000, including the Islamic Media Center, azzam.com and maalemaljihad.com (the first al-Qaeda website, created in 2000) (Hecker 2015), closely followed by alneda.com (Stenersen 2008). A few years later, webforums started to emerge and became the main location for online meetings and jihadist hubs. Popular at the time and somewhat a forerunner to social media, webforums provided a place for sharing ideas and for questions and answers. Also, those forums circulated tutorials teaching various modi operandi as well as instructions on how to build or use weapons (Zelin 2013). As early as 2008, some jihadists active in the al-Fallujah Forum were advocating the leveraging of mainstream social media or sharing platforms such as Facebook or YouTube (al-Shishani 2010).

However, social media use became common only a few years later, aided by a generation of foreign terrorist fighters who joined the Syrian civil war from 2011. Journalist David Thomson referred to them as the 'Sheikh Google and LOL jihad' (Thomson 2014).

This phenomenon was more openly acknowledged in 2015, when numerous accounts and profiles, in particular on Twitter, were deleted by the platform. Indeed, social media companies, under political, legal and ethical pressure in the wake of the Paris terror attacks,[1] started to intervene more, with a clear consequence on account and content reach and longevity. Research has highlighted that accounts linked to Islamic State (IS) were subject to more frequent and faster deletions than accounts linked to other jihadist groups (Conway et al. 2017). IS and other terror groups, seeking spaces less likely to be disrupted by authorities or by the platforms themselves, migrated towards Telegram, the encrypted messaging application. They chose relative operational security and longevity at the price of a more restricted audience than they would have on social media (Bindner and Gluck 2017).

After the crackdown on jihadist use of social media platforms, ISIS and others were quickly followed to Telegram by al-Qaeda. Telegram combines three key functionalities. First, it is an encrypted messaging application tool. This functionality may be used in one-to-one conversations, in groups or in channels. Channels are especially favoured for their unidirectional functionality, which allows users to broadcast messages to an audience. Second, Telegram groups have functions similar to social media and webforums, offering interactive and multidirectional communication. And third, it is used as a hosting platform, where channel administrators or group members may post all types of media (including large files of 1.5 GB) and documents that can be uploaded in Telegram, aggregating in that regard a significant jihadi library of new and old or recycled materials (Clifford and Powell 2019).

Telegram thus constitutes a secure support system, and jihadis and their supporters use it as a launching pad to propel their material to other platforms and social media with numerous 'salvoes' of URLs. For instance, a recent analysis assessed the dissemination extent and life expectancy of these URLs pointing to an Amaq video and

1 Specifically, Charlie Hebdo attack on 7 January, 2015, Montrouge killing on 8 January and the HyperCasher attack on 9 January.

one issue of IS's weekly al-Naba', totalling 60 URLs pointing to 20 platforms for the former and 88 URLs pointing to 19 platforms for the latter (Bindner and Gluck 2018). These links are immediately published and shared on mainstream social media, thus targeting a significant pool of potential new recruits. This emergence on the surface web essentially aims to infuse the message into social media, and thus to broaden the communication and recruitment reach to a wider audience. This audience can later be funnelled back to Telegram, where exposure to propaganda turns out to be more intense. This phenomenon takes place at a continuous pace, repeating itself each time a new piece of propaganda content is released, thus being a continuous and sustained back-and-forth movement between the deep web and the surface web. A simple photo essay showing luscious fruit ready for harvest or freshly baked goods for sale in an area controlled by IS could be uploaded online to a Telegram channel with an instruction to spread these photos via social media accounts as a form of PR. Social media accounts created to display these 'utopic' photo essays sometimes contain links to more egregious content hidden in Telegram channels. One significant characteristic of this content is its multifaceted segmentation, with the aim to reach as large an audience as possible. The case in point of this segmentation is the variety of Telegram channels and groups that funnel content towards specific target audiences. IS has, for instance, multiple media outlets to translate its news publications in dozens of languages, as well as recycling old content. It also has groups segmented by gender, tech channels dedicated to those interested in anonymity, and channels dedicated to news, and others to theory and dogma.

Within social media, tactics used by jihadist extremists evolve continuously to circumvent monitoring and enable content to emerge despite the increasing implementation of artificial intelligence (AI). A simple example would be image distortion or cropping (logos) to cheat AI (Atiyas Lvovsky 2017), hijacking hashtags, posting egregious content in comments instead of main posts to avoid detection, or using profile pics, bios or names often not openly suggesting a link to the jihadist movement. Branding an account is also a recurring tactic, in which a user makes repeated use of the same profile picture

(or also the same ID, background picture or self-description) to create recognition that is sometimes so powerful that newly created social media accounts are able to attract dozens of followers or friend requests within minutes (Bindner and Gluck 2017). Other recent trends include 'hiding' media in multiple uploads in a single post, 'blank screen' previews of videos and the increasing use of disappearing stories. Jihadist circumvention of monitoring even extends to Telegram monitoring as well as reporting on content that led to a rise in deletions of channels and groups. For example, an Islamic State of Iraq and Syria (ISIS) watch bot created by Telegram logs the number of ISIS-linked channels deleted on a daily basis. Tactics on Telegram include the use of buttons to join channels and groups, in addition to hyperlinks, changing channel and group logos frequently, and using special fonts for usernames. Suspicion towards the app from its extremist users has increased over the past year, primarily when Telegram CEO Pavel Durov announced in August 2018 that the platform might decide to provide IP addresses or phone numbers to governments when cases are linked to terrorism.[2] For years now, channels have been mostly private, meaning that they can only be joined with a key-like URL, obtained in other channels or groups, or upon direct request, and sometimes cloaked as buttons (with no way of sharing other than in intended channels or groups). Some groups may only be joined upon request. Names of channels may be distorted to avoid automatic detection. The use of bots to provide new links to channels has become common, as well as the use of URL shorteners (named 'permanent' or 'eternal' links) to content or channels, automatically redirecting to the latest available page, following deletion of a previous link.

If Telegram still nevertheless remains the preferred platform for jihadists to date (Clifford and Powell 2019), the dissemination scheme has evolved during the past year, a consequence of tighter monitoring from social media, where content emerges on a smaller

2 www.cbsnews.com/news/telegram-encrypted-messaging-app-cooperate-terror-investigations-not-russia/

scale and activists often maintain a lower profile. Their weaker presence on social media is widely offset by a combination of presence on other parts of the surface web. To share content, jihadi activists have diversified their outlets, testing new apps and disseminating their material on alternative platforms. In recent months, official ISIS content has emerged on other social networks or messaging apps such as Yahoo Together, Viber, TamTam, Baaz and Zello, and on various decentralised platforms (Bodo 2018), such as Riot, Minds and Rocket. Chat, and more recently even on Mastodon and Gab.ai, to name but a few. To host content, jihadists now exploit various platforms, be it cloud-based storage services – such as Google's, Amazon's, OneDrive and so on – with key-like links widely shared, and multiple links pointing to cloud storage links for each release, or more 'obscure platforms', which are unknown to most but are still within easy reach via search engines.

The fragmentation of propaganda across multiple platforms has several consequences in terms of countering online jihadism. First, increased monitoring on mainstream social media curbed the viral effect of this material and isolated it from the public to a certain extent. Second, this recent dissemination pattern makes it less traceable by authorities, who end up playing a game of 'cat and mouse' but running in all directions. This poses the risk of a more diffuse online presence, which raises the question of platforms' liability regarding published content, and the relevance and conditions of their monitoring and censorship.

Nevertheless, jihadi media activists are still present on these platforms, either in the form of 'dormant' cells, sharing various types of non-violent content but reaching potential recruits or sympathisers on private messages mostly on alternative secured platforms, or with accounts that are doomed to be short-lived, used to make specific elements of propaganda emerge. As illustrated in the al-Hayat Media Center video 'Inside the Khilafah n.8', released on 30 October, 2018, online *munasireen* (sympathisers) were encouraged to remain active and resilient: "Strive patiently in the digital arena, and do not allow

the disbelievers to enjoy a moment of sleep or to live a pleasant life: if they close one account, open another three, and if they close three, open another thirty."

It is in this context that the European Commission is about to adopt new regulations[3] modifying the current doctrine of platforms' liability. According to the current doctrine, based on the 2000 E-Commerce Directive, platforms benefit from a liability exemption regarding illicit content they host.[4] This proposed Regulation, despite not disputing the principles of the 2000/31/EC E-Commerce directive[5] and the liability exemption, constitutes an additional step towards a voluntary approach – both proactive (active monitoring is encouraged, and some content deletion will be performed on a discretionary stance) and reactive (with a very short time frame of one hour granted to the hosting platform for content removal). This new regulation will apply to a significant range of platforms, even small actors, as long as they have commercial activity in the EU. Whereas continuous pressure against the emergence of terror content is necessary for political, ethical and social reasons, this raises the issue of a form of externalisation of content monitoring – traditionally a government's prerogative – towards the private sector. Moreover, fines – with a deterrence objective – that may apply to platforms that regularly fail to comply may reach significant amounts (up to 4% of the company global revenue). Financial pressure is thus applied on the sector, which will most certainly disrupt small platforms' viability. To avoid those fines and remain on the 'safe side', excessive monitoring might then take place *de facto* and affect the current stance on freedom of speech.

3 See www.eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018PC0640&from=FR

4 Provided the content was uploaded without their knowledge and provided they remove it as soon as they have the knowledge of it.

5 See www.eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32000L0031

This variety, agility and enduring presence of jihadi extremists online raises trade-offs the authorities are confronted with. Account deletion cannot be a definitive solution, given activists' resilience, and given the fact that open-source intelligence and social media intelligence now constitute an important source for the authorities. Moreover, which entity should be responsible for such a monitoring? Is this a state prerogative or a private sector one? What type of content should be taken down and which party should draw a line between 'acceptable' and 'unacceptable' content? How to set a balance between human reviewers on the one side, and AI and algrithms on the other side? Besides all these issues also lie the question of digital evidence and the necessity to maintain a fluid relationship between private actors who own this data and the authorities who require them, in particular to feed legal cases.

References

Al-Shishani, M.B. 2010. 'Taking al-Qaeda's Jihad to Facebook'. *Terrorism Monitor*, 8 (5). Retrieved from: www.jamestown.org/program/taking-al-qaedas-jihad-to-facebook/

Atiyas Lvovsky, L. 2017. 'Imposters and Deception on the Internet'. International Institute for Counter-Terrorism. Retrieved from: www.ict.org.il/Article/2163/terrorism-impostes-and-deception-on-the-internet

Berger, J.M. and Morgan, J. 2015. 'The ISIS Twitter Census: Defining and Describing the Population of ISIS Supporters on Twitter'. Brookings. Retrieved from: www.brookings.edu/wp-content/uploads/2016/06/isis_twitter_census_berger_morgan.pdf

Bindner, L. and Gluck, R. 2017. 'Wilayat Internet: ISIS' Resilience across the Internet and Social Media'. *Bellingcat*. Retrieved from: www.bellingcat.com/news/mena/2017/09/01/wilayat-internet-isis-resilience-across-internet-social-media/

Bindner, L. and Gluck, R. 2018. 'Trends in Islamic State's Online Propaganda: Shorter Longevity, Wider Dissemination of Content'. International Centre for Counter-Terrorism – The Hague. Retrieved from: www.icct.nl/publication/trends-in-islamic-states-online-propaganda-shorter-longevity-wider-dissemination-of-content/

Bodo, L. 2018. 'Decentralized Terrorism: The Next Big Step for the So-Called Islamic State (IS)?'. *VOX-Pol Network of Excellence*. Retrieved from: www.voxpol.eu/decentralised-terrorism-the-next-big-step-for-the-so-called-islamic-state-is/

Clifford, B. and Powell, H. 2019. 'Encrypted Extremism: Inside the English-Speaking Islamic State Ecosystem on Telegram'. The Georges Washington University. Retrieved from: www.extremism.gwu.edu/sites/g/files/zaxdzs2191/f/EncryptedExtremism.pdf

Conway, M., Khawaja, M., Lakhani, S., Reffin, J., Robertson, A. and Weir, D. 2017. 'Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts'. *VOX-Pol Network of Excellence*.

Hecker, M. 2015. 'Websocial et djihadisme'. *Focus stratégique*, n.57, Centre des études de sécurité de l'IFRI, p. 12. Retrieved from: www.ifri.org/sites/default/files/atoms/files/fs57hecker.pdf

Lia, B. 2008. 'Architect of Global Jihad: The Life of Al-Qaeda Strategist Abu Mus'ab Al-Suri'. Columbia University Press.

Stenersen, A. 2008. 'The Internet: A Virtual Training Camp'. Terrorism and Political Violence, 20 (2), pp. 215–233.

Thomson, D. 2014. *Les Français jihadistes*. Les Arènes.

Zelin, A.Y. 2013. 'The State of Global Jihad Online: A Qualitative, Quantitative, and Cross-Lingual Analysis'. New America Foundation. Retrieved from: www.washingtoninstitute.org /uploads/Documents/opeds/Zelin20130201-NewAmericaFoundation.pdf

RIGHT-WING EXTREME DIGITAL SPEECH IN EUROPE AND NORTH AMERICA

Bharath Ganesh

Right-wing extreme digital speech (RWEDS) has become a key concern in recent years due to the proliferation of hate speech, racism and white nationalist rhetoric on social media platforms. Research into RWEDS suggests that the communications platforms that right-wing extremists use are important to forging a transnational subculture grounded in racism, hate and white nationalism (Back, Keith and Solomos 1998, in Perry and Olsson 2009, p. 190). Technical changes in digital communication have reconfigured the ways in which this global racist subculture operates. After explaining the definitional challenges in RWEDS, this essay explores how these technical configurations have been used by right-wing extremists.

DEFINITIONAL ISSUES IN RIGHT-WING EXTREME DIGITAL SPEECH

Drawing the boundaries on what counts as RWEDS is extremely difficult. This issue is less evident when we focus specifically on violent extremists. However, when we turn our attention to extreme digital speech, the boundaries of belonging to a specific group, organisation, movement or party are blurred. These blurry boundaries between far-right politics, social movements and hateful subcultures complicate the differentiation of legitimate political opinion from extreme digital speech.

Particularly in the social media context, RWEDS is often expressed as 'hate speech', which is a highly contested term in itself. Hate speech has similarities with RWEDS insofar as both share a double function: to 'dehumanize and diminish members' of a targeted group while simultaneously signalling 'to let others with similar views know they are not alone [and] to reinforce a sense of

an in-group that is (purportedly) under threat' (Gagliardone et al. 2015, pp. 10–11). This definition of hate speech proposed for the online context is also particularly relevant to understanding right-wing extreme digital speech. In Extremism, J.M. Berger writes that an extremist ideology can be understood 'as a collection of texts' that define the in-group, the out-group and their interactions with one another (2018a, p. 26). In RWEDS in Europe and North America, the white population is rendered as the in-group while the out-group has shifted in different contexts. Nativism, which is the belief that a state should be inhabited only by those constructed as 'native' to its territory, gives consistency to different instances of RWEDS across Europe and North America (see Mudde 2007, p. 22). Building on this concept, Froio argues that nativism, which incorporates both racist arguments as well as neo-racist arguments that focus on cultural (rather than ethnic or racial) differences, is also a core feature of the digital communication of far-right social movements in France (2018, p. 698). Nativism has also been shown to be central to the transnational linkages and the global racist subculture that RWEDS is involved in producing (cf. Caiani and Kröll 2015; Froio and Ganesh 2018).

Like RWEDS, definitions of the far right also suffer from complexities in deciding what types of individuals, social movements and political parties ought to be included under its umbrella. In their description of the far right as a 'galaxy of *nativist* actors including extreme and radical right organisations', Castelli Gattinara and Pirro (2018, p. 4, emphasis added) provide a particularly apt formulation for thinking about nativism as the common thread that connects the heterogeneous networks that house and foster RWEDS.

THREE CONFIGURATIONS OF THE GALAXY OF NATIVIST ACTORS: THE INTERNET FORUM, THE BLOG AND THE SWARM

This section considers three configurations of RWEDS from the early 1980s to the late 2010s. While these configurations are outlined in chronological order to ease comprehension, it is important to remember that all three of these configurations are present in the contemporary networks of RWEDS.

Internet Forums and the Virtual Extremist Community

The earliest documentation of RWEDS was of white supremacist, racist and neo-Nazi bulletin board systems that began to emerge in the 1980s. Berlet (2001) identifies 1984 as the year 'when hate went online', pointing to early bulletin board systems (BBSs) such as 'Aryan Liberty Net' and 'White Aryan Resistance BBS'. BBSs were among the first connective systems used to develop a global racist subculture. Later, graphical user interfaces (GUIs) for web browsing began to emerge, and modems with sufficient bandwidth became more widely available. Therefore, BBSs began to fade in their prominence and were increasingly replaced by Internet forums, or webforums, that could be accessed with graphical browsers.

Right-wing extremism has a long history online (Back 2002; Burris, Smith and Strahm 2000). For example, Usenet forums were one of the first virtual communities of right-wing extremists, where white supremacists, neo-Nazis and even members of the Ku Klux Klan would connect with one another (Daniels 2009; Zickmund 2002). In 1995, Don Black started the best-known extreme right-wing website, Stormfront.org, which describes itself as a home for the 'embattled White minority' and evolved into an interactive forum. The forum – which hosts white supremacists, Christian fundamentalists, neo-Nazis and Holocaust deniers (to name just a few of the extreme ideologies professed on Stormfront) – has been frequently described as a virtual community (Bowman-Grieve 2009; De Koster and Houtman 2008). For participants on Stormfront, the forum enhances a sense

of community as members 'encourage each other to express opinions and viewpoints on topics of interest, and the eager challenging of each other's perspectives' (Bowman-Grieve 2009, p. 997). Stormfront has also served as a 'safe space' for articulating extreme opinions that are often considered unacceptable in other online and offline communities. It thereby gives its members a sense of a 'second home', deepening emotional attachments between users (De Koster and Houtman 2008). The forum allows for a persistent, searchable site in which RWEDS could safely be expressed. While there is little evidence to show that places like Stormfront were central to the planning and execution of attacks, they are key sites for the spread of extreme right-wing ideology. As an analysis of Anders Breivik's activity on the forum shows, extreme virtual communities provide resources for potential extremists that facilitate radicalisation and the use of violence (Gardell 2014; Ravndal 2013). Furthermore, these websites have been found to be a crucial part of the enculturation into an extreme worldview that gives would-be terrorists a narrative that justifies their actions (Holt, Freilich and Chermak 2016).

The Blog and the Counter-Jihad

By the early 2000s, many extremists were setting up websites that were both covertly and overtly racist, often in the form of blogs. The blog reduced the barriers to entry for individuals who could not easily publish their views and disseminate them to a broad audience, allowing them to circumvent gatekeepers in the media and to fashion themselves into far-right thought leaders. Similar tendencies are present on YouTube, where vlogs (video blogs) offered the opportunity for individuals to position themselves as pundits or moral entrepreneurs with access to potentially massive audiences. The rise of the blog was particularly beneficial for the counter-jihad movement that was emerging in the mid-2000s in Europe and North America (Ekman 2015). Blogs were established by prominent individuals in the counter-jihad movement, a transnational movement of (among other types of actors) political bloggers, think tanks, street movements and campaign groups (see Ekman 2015; Lee 2016; Froio 2018) that have

used Web 2.0 technologies effectively to spread their messages and connect like-minded groups across Europe and North America. Lee (2016) notes that the sections of the counter-jihad scene can be differentiated from the 'extreme' versus the 'merely hawkish' because of their use of conspiracy theories about 'Islamisation' and the treatment of Islam and Muslims as a homogeneous bloc (2016, pp. 259–260). In producing these websites and often focusing on inventing the 'threat' presented by Islam and Muslims to Western society (rather than engaging in obvious racist discourse, as is more common on webforums), these sites pass themselves off as legitimate journalism. This makes the differentiation of RWEDS from legitimate political opinion even more blurry, despite many of these blogs – such as *Jihad Watch* (Robert Spencer), the *Geller Report* and *Atlas Shrugs* (Pamela Geller) and *Gates of Vienna* – mobilising conspiracy theory and anti-Muslim hate (cf. Allen 2013; Ekman 2015). After 9/11 much of the attention of the extreme right turned towards Muslims and migration into Europe from the Islamic world (Zúquete 2008). Today, 'Islamisation' and the fear of Muslims taking over Western countries have become a driving force for many of the policy positions of the extreme right and their overlap with existing far-right political parties, with the blogs mentioned above playing a prominent role (Southern Poverty Law Center n.d.; Ekman 2015). While these blogs do not advocate violence against Muslims, they have been crucial in providing information in highly biased forms that contribute to the narratives used in extreme digital speech (Lee 2016). Political blogs associated with the counter-jihad movement linked with and contributed to protests, such as that against the Park51 mosque in New York, and inspired movements such as defence leagues across northern Europe (see BBC News 2013; Meleagrou-Hitchens and Brun 2013; Ekman 2015).

The political blog, particularly in the case of the counter-jihad, played an important role in the collection of texts that advance an 'in-group' under threat from an 'out-group', rendered in these cases as Muslims. These blogs often curated the content of different bloggers in the counter-jihad scene, though these bloggers' extreme digital speech is not likely to be a central focus of any average member

of these defence leagues (see Lee 2015, p. 265). Rather, extreme digital speech in the form of blogging in the counter-jihad nebula is more likely to be a 'tool for opinion leaders', in which their curation and production of content (and links to the media) serve to reinforce and support the core beliefs of viewers. In this sense, the extreme digital speech of actors in the counter-jihad scene might situate them as "the unofficial [custodians] of the counter-jihad discourse, supporting it where they can and protecting it from criticism where necessary" (2015, pp. 264–265). Thus the configuration of blogs has a particular purpose not of simply using RWEDS to produce an echo chamber, but form a 'patchwork' of ideas and texts that reinforce a nativist worldview in which Muslims are seen as a monolithic threat to the public in Europe and North America.

The Swarm, Social Media and the Alt-Right

Whereas the webforum represents a technical configuration that produced virtual communities, and the political blog allowed for a decentred curation of information that informed the counter-jihad 'nebula', social media represents a third technical configuration. The webforum afforded a relatively closed online community of like-minded people that one would have to join, but social media facilitates public spheres in which subcultures can communicate with one another. Thus, it became possible for networked action between proponents of RWEDS to advance their dissemination of narratives to a broader audience. Social media is configured in such a way that it allows a small fringe of extremists to spread their narratives and amplify the voices of the bloggers, video 'journalists' and pundits whom they prefer across a much larger audience than could be afforded in a webforum. Furthermore, social media's interactive features not only enable the amplification of extremist voices, but also enable users to engage directly in an online subculture.

The metaphor of the swarm is perhaps the most useful way of trying to define the networks of right-wing extreme digital speech (cf. Ganesh 2018). Social media has enabled these networks to quickly amplify certain narratives, harass and attack their

opponents, and influence mainstream discussion (Marwick 2018). Elements of the webforum configuration are still present in places such as 4chan and Reddit, which are sites in which relatively small communities produce memes, images, video and other content and co-ordinate activities on mainstream social media platforms such as Twitter and Facebook (cf. Zannettou et al. 2018; Topinka 2017; Ludemann 2018; Massanari 2017). From there, users co-ordinate what are referred to as 'raids' on platforms, in which they co-ordinate the dissemination of memes and other content with the explicit aim of affecting public opinion.

Unlike jihadism, the exponents of RWEDS occupy a legal grey area that makes policing them much different. This swarm is trans-national, and relations can be drawn across the world, but in Europe and North America, there are two major umbrella groups that need to be addressed. The first is a primarily Anglophone network often referred to as the 'alt-right' (Davey and Ebner 2017; Hawley 2017; Nagle 2017; Salazar 2018), which has grown out of Internet trolling cultures and white nationalist organising (Phillips 2018). The alt-right often draws on a swarm that incorporates users from the 'manosphere' and misogynistic online subcultures (Ging 2017; Marwick and Caplan 2018). Much of their content is highly emotive and associated with a reactionary white identity politics (Pollard 2018). Many of the core myths of the alt-right are shared by Identitarians in Western Europe, which have close links with alt-right networks and an intellectual lineage based on the writings of the *Nouvelle Droite* (Bar-On 2018).

These groups are often explicitly non-violent in their public statements both online and in person. This is true, for example, of the group Britain First, which had amassed over a million likes on its Facebook page. However, violent and hateful statements, as well as antagonism towards Muslims, are common amongst the many followers of Britain First and other nativist groups that have used digital communication in Europe (for example, Faith Matters 2014; Evolvi 2017; in Sweden see Törnberg & Törnberg 2016). What the swarm metaphor indicates is that the blurry boundaries between groups can allow them to exist as part of an 'extremist bloc' which is indicative of a tenuous 'coalition' of activists online

(see Berger 2018b, p. 53) and is also evident amongst the counter-jihad (see Önnerfors 2017). Thus, rather than consider the alt-right alone, the term 'swarm' turns our attention to a *coalition* of various toxic cultures, including white nationalists, counter-jihad activists and misogynists (Ganesh 2018; see also Ging 2017; Massanari 2017). Some members of this swarm (among many other actors who might identify differently) use digital communication to foster the enculturation of audiences into an extreme worldview in which the global 'white race' is under attack by liberal democratic norms of pluralism and diversity (Berger 2016; Daniels 2018; Dentice and Bugg 2016; Gardell 2014). This swarm relies on long-established narratives of the 'white race' under attack, similar to white nationalist and white supremacist language (Atkinson 2018; Bjork-James and Maskovsky 2017; Doerr 2017; Eckhouse 2018).

Providing evidence that an RWEDS swarm is growing on social media platforms is difficult because open historical datasets are not freely available to researchers. However, there is evidence of significant growth. For example, J.M. Berger identifies that followership of white nationalist accounts grew from 3,542 in 2012 to 25,406 in 2016 (Berger 2016, p. 4). In his recent *Alt-Right Twitter Census* (2018b), Berger writes that those associated with the alt-right on Twitter likely exceeds 200,000 users, which is presented as a highly conservative estimate. While the two studies use rather different measures (the first explores a number of white nationalist accounts, whereas the latter takes a broader view on followers), Berger's two studies seem to suggest an increase in the number of 'alt-right', white nationalist, far-right and Identitarian accounts in the past six years.

EMERGING CHALLENGES IN COUNTERING RWEDS

As Reddit, Facebook and Twitter have increasingly cracked down on those expressing RWEDS for violating hate speech and harassing users, alternative websites have been set up specifically for the use of RWEDS. Gab, for example, is a 'free speech' alternative to Twitter that hosts a multitude of extreme right-wing users that frequently

use racist and anti-Semitic language (Zannettou et al. 2017; Davey and Ebner 2017). These are places where hateful narratives that can potentially inspire hate crimes, and violent extremism can be accessed. This makes their governance highly controversial and very difficult. Furthermore, any crackdown on extreme right-wing accounts can result in a backlash centred around claims of 'political correctness' and 'free speech' that reproduce 'evidence' of the mythical 'liberal-multicultural elite' attacking white voices.

Looking at the main sites of communication for the swarm provides a sense of how the new media is used by these groups. Generally, many of the activists on the alt-right from across the world have large YouTube followings (such as Lauren Southern, Richard Spencer's 'AltRight.com' channel and Red Ice TV). Further, less prominent activists seeking more attention from colleagues in the alt-right movement are using YouTube to gain notoriety and audiences. Their content is widely distributed on Facebook, Twitter and Reddit to broad audiences (which cannot easily be quantified). Where exponents of RWEDS face disruption from these platforms, they tend to migrate to other social media platforms that have less strict regulations on hate speech and content regulation, though they are less desirable for exponents of RWEDS because of their much smaller audiences.

RWEDS is a growing concern, and the relationship between the Internet and recent attacks by right-wing extremists have shown that RWEDS can motivate extreme right-wing terrorism, particularly by providing a narrative and community that facilitates radicalisation. However, RWEDS is neither a necessary nor sufficient condition for the execution of a violent act. Rather, RWEDS potentially enculturates audiences into an exclusionary worldview in which violence against minorities can be justified. This presents a significant problem for existing laws pertaining to countering extremism and hate speech, which often require specific calls for violence against protected groups or the facilitation of terrorist attacks online. This legal grey area that they operate in allows this swarm to avoid prosecution while spreading extremist messaging online.

References

Allen, D.C. 2013. *Islamophobia*. Ashgate Publishing, Ltd.

Atkinson, D.C. 2018. 'Charlottesville and the Alt-Right: A Turning Point?' *Politics, Groups, and Identities*, 6 (2), pp. 309–315.

Back, L. 2002. 'Aryans Reading Adorno: Cyber-culture and Twenty-First Century Racism'. *Ethnic and Racial Studies*, 25 (4), pp. 628–651.

Bar-On, T. 2018. 'The Radical Right and Nationalism'. In J. Rydgren (Ed.), *The Oxford Handbook of the Radical Right* (pp. 17–41). Oxford: Oxford University Press.

BBC News 2013. 'US Bloggers Banned from Entering UK'. *BBC News*, 26 June. Retrieved from: www.bbc.co.uk/news/uk-23064355

Berger, J.M. 2016. 'Nazis vs. ISIS on Twitter: A Comparative Study of White Nationalist and ISIS Online Social Media Networks (Program on Extremism Occasional Paper)'. George Washington University. Retrieved from: www.extremism.gwu.edu/sites/g/files/zaxdzs2191/f/ downloads/Nazis%20v.%20ISIS.pdf

Berger, J.M. 2018a. *Extremism*. 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*.

Berlet, C. 2001, 'When Hate Went Online'. Northeast Sociological Association Spring Conference in April, pp. 1–20.

Bjork-James, S. and Maskovsky, J. 2017. 'When White Nationalism Became Popular'. *Anthropology News*, 58 (3), e86–e91.

Bowman-Grieve, L. 2009. 'Exploring "Stormfront": A Virtual Community of the Radical Right'. *Studies in Conflict & Terrorism*, 32 (11), pp. 989–1007.

Burris, V., Smith, E. and Strahm, A. 2000. 'White Supremacist Networks on the Internet'. *Sociological Focus*, 33 (2), pp. 215–235.

Caiani, M. and Kröll, P. 2015. 'The Transnationalization of the Extreme Right and the Use of the Internet'. *International Journal of Comparative and Applied Criminal Justice*, 39 (4), pp. 331–351.

Castelli Gattinara, P.C. and Pirro, A.L.P. 2018. 'The Far Right as Social Movement'. European Societies, 21 (4), pp. 447–462.

Council on American-Islamic Relations [CAIR]. (2017). 'Pamela Geller'. *Counter-Islamophobia Project*. Retrieved from: www.islamophobia.org/islamophobic-individuals/pamela-geller/77-pamela-geller.html

Daniels, J. 2009. *Cyber Racism: White Supremacy Online and the New Attack on Civil Rights*. Rowman & Littlefield Publishers.

Daniels, J. 2018. 'The Algorithmic Rise of the "Alt-Right"'. *Contexts*, 17 (1), pp. 60–65.

Davey, J. and Ebner, J. 2017. *The Fringe Insurgency* Institute for Strategic Dialogue. Retrieved from: www.isdglobal.org/wp-content/uploads/2017/10/The-Fringe-Insurgency-221017.pdf

De Koster, W. and Houtman, D. 2008. 'Stormfront is like a Second Home to Me'. *Information, Communication & Society*, 11 (8), pp. 1,155–1,176.

Dentice, D. and Bugg, D. 2016. 'Fighting for the Right to Be White: A Case Study in White Racial Identity'. *Journal of Hate Studies*, 12 (1), p. 101.

Doerr, N. 2017. 'Bridging Language Barriers, Bonding against Immigrants: a Visual Case Study of Transnational Network Publics Created by Far-Right Activists in Europe'. *Discourse & Society*, 28 (1), pp. 3–23.

Eckhouse, L. 2018. 'White Riot: Race, Institutions, and the 2016 U.S. Election'. *Politics, Groups, and Identities*, pp. 1–12.

Ekman, M. 2015. 'Online Islamophobia and the Politics of Fear: Manufacturing the Green Scare'. *Ethnic and Racial Studies,* 38 (11), pp. 1,986–2,002.

Evolvi, G. 2017. '#Islamexit: Inter-Group Antagonism on Twitter'. *Information, Communication & Society*, 22 (3), pp. 386–401.

Faith Matters 2014. *Facebook Report: Rotherham, Hate and the Far-Tight Online*. Faith Matters. Retrieved 4 December, 2018 from: www.tellmamauk.org/rotherham-hate-and-the-far-right-online/

Froio, C. 2018. 'Race, Religion, or Culture? Framing Islam between Racism and Neo-Racism in the Online Network of the French Far Right'. *Perspectives on Politics,* 16 (3), pp. 696–709.

Froio, C. and Ganesh, B. 2018. 'The Transnationalisation of Far Right Discourse on Twitter'. *European Societies*, 21 (4), pp. 513–539.

Gagliardone, I., Gal, D., Alves, T., and Martinez, G. 2015. *Countering Online Hate Speech*. UNESCO Publishing.

Ganesh, B. 2018 'The Ungovernability of Digital Hate Culture'. *Journal of International Affairs*, 71 (2), pp. 30–49.

Gardell, M. 2014. 'Crusader Dreams: Oslo 22/7, Islamophobia, and the Quest for a Monocultural Europe'. *Terrorism and Political Violence*, 26 (1), pp. 129–155.

Ging, D. 2017. 'Alphas, Betas, and Incels: Theorizing the Masculinities of the Manosphere'. *Men and Masculinities*. 1097184X17706401.

Hawley, G. 2017. Making Sense of the *Alt-Right*. Columbia University Press.

Holt, T.J., Freilich, J.D. and Chermak, S.M. 2016. 'Internet-Based Radicalization as Enculturation to Violent Deviant Subcultures'. *Deviant Behavior*, 38, pp. 855–869.

Lee, B. 2015. 'A Day in the "Swamp": Understanding Discourse in the Online Counter-Jihad Nebula'. *Democracy and Security*, 11 (3), pp. 248–274.

Lee, B. 2016. 'Why We Fight: Understanding the Counter-Jihad Movement'. *Religion Compass,* 10 (10), pp. 257–265.

Ludemann, D. 2018. '/pol/emics: Ambiguity, Scales, and Digital Discourse on 4chan'. *Discourse, Context & Media*, 24, pp. 92–98.

Marwick, A.E. 2018. 'Why Do People Share Fake News? A Sociotechnical Model of Media Effects'. *Georgetown Law Technology Review*, 2 (2), pp. 474–512.

Marwick, A.E. and Caplan, R. 2018. 'Drinking Male Tears: Language, the Manosphere, and Networked Harassment'. *Feminist Media Studies*, 18 (4), pp. 543–559.

Massanari, A. 2017. '#Gamergate and The Fappening: How Reddit's Algorithm, Governance, and Culture Support Toxic Technocultures'. *New Media & Society*, 19 (3), pp. 329–346.

Meleagrou-Hitchens, A. and Brun, H. 2013. 'A Neo-Nationalist Network: The English Defence League and Europe's Counter-Jihad Movement'. ICSR: Kings College London.

Mudde, C. 2007. *Populist Radical Right Parties in Europe*. Cambridge University Press.

Nagle, A. 2017. *Kill All Normies: Online Culture Wars from 4chan and Tumblr to Trump and the Alt-Right*. John Hunt Publishing.

Önnerfors, A. 2017. 'Between Breivik and PEGIDA: The Absence of Ideologues and Leaders on the Contemporary European Far Right'. *Patterns of Prejudice*, 51 (2), pp. 159–175.

Perry, B. and Olsson, P. 2009. 'Cyberhate: The Globalization of Hate'. *Information & Communications Technology Law*, 18 (2), pp. 185–199.

Phillips, W. 2018. *The Oxygen of Amplification* (Part 1). New York: Data & Society. Retrieved from: www.datasociety.net/output/oxygen-of-amplification/

Pollard, T. 2018. 'Alt-Right Transgressions in the Age of Trump'. *Perspectives on Global Development and Technology*, 17 (1–2), pp. 76–88.

Ravndal, J.A. 2013. 'Anders Behring Breivik's Use of the Internet and Social Media'. *Journal Exit-Deutschland. Zeitschrift Für Deradikalisierung Und Demokratische Kultur*, ss2, pp. 172–185.

Salazar, P.J. 2018. 'The Alt-Right as a Community of Discourse'. *Javnost – The Public*, 25 (1–2), pp. 135–143.

Southern Poverty Law Center. n.d. 'Pamela Geller'. Retrieved from: www.splcenter.org/fighting-hate/extremist-files/individual/pamela-geller [Accessed on 17 August 2018].

Topinka, R. J. (2017). 'Politically Incorrect Participatory Media: Racist Nationalism on r/ImGoingToHellForThis'. *New Media & Society*, 1461444817712516.

Törnberg, A. and Törnberg, P. 2016. 'Muslims in Social Media Discourse: Combining Topic Modeling and Critical Discourse Analysis'. *Discourse, Context & Media*, 13, pp. 132–142.

Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringhini, G. and Blackburn, J. 2018. 'What is Gab? A Bastion of Free Speech or an Alt-Right Echo Chamber?' ArXiv:1802.05287 [Cs].

Zannettou, S., Caulfield, T., De Cristofaro, E., Kourtelris, N., Leontiadis, I., Sirivianos, M., Stringhini, G. and Blackburn, J. 2017. 'The Web Centipede: Understanding How Web Communities Influence Each Other through the Lens of Mainstream and Alternative News Sources'. *Proceedings of the 2017 Internet Measurement Conference* (pp. 405–417). New York, NY, USA: ACM.

Zickmund, S. 2002. 'Approaching the Radical Other: The Discursive Culture of Cyberhate'. In S. Jones (Ed.), *Virtual Culture: Identity and Communication in Cybersociety* (pp. 185–206). London: SAGE.

Zúquete, J. (2008). 'The European Extreme-Right and Islam: New Directions?' *Journal of Political Ideologies*, 13 (3), pp. 321–344.

IMPACT OF CONTENT

Mubaraz Ahmed

Radicalisation remains a contentious area of study because the role of online interactions and propaganda on radicalisation processes remains a 'highly contested subject' (Meleagrou-Hitchens and Kaderbhai 2017, p. 6). This debate involves what constitutes radicalisation and which type of Internet activity (propaganda, interactions and so on) should fall into this category. While academic discourse on the matter is contentious, unfortunately public discourse is dramatic. The role of extremist content is often sensationalised, with reports of Internet use by violent extremists used as fodder for front-page headlines blaming tech platforms for being complicit in or indifferent to extremist violence. Statements from senior politicians and police officials blaming tech companies for abuses of their platforms also contribute to the increasingly adversarial public discourse on this topic.

As a result, policymaking in this space is increasingly based on the dramatic combination of public pressure and unproven assumptions shaped by anecdotal evidence rather than rigorous, empirical research. Recent research efforts have sought to address this gap by quantifying the nature of Internet use among terrorist actors and exploring content types consumed by individuals convicted of terrorist activity. In view of the contention and research gaps on online radicalisation, this essay first reviews the current academic landscape on the relationship between online and offline extremist activity. It then examines case studies of violent extremist perpetrators to explore the nature and type of Internet activity prior to engaging in violence. And finally, it looks ahead to what is required from a policy and research perspective to better comprehend and address the impact and role of extreme digital speech.

The following paragraphs will explore some of the prominent ideas and perspectives across different sectors that have emerged from research into the interaction between online engagement and offline behaviour. And they will examine what role the Internet plays in violent extremist activity.

Extremist groups use the Internet in three primary ways: to circulate ideological messaging; enable users to participate discreetly; and facilitate engagement with a social environment. The ease of access, low cost and reach of the Internet means that the threshold for entry into extremist networks is lowered, while the privacy and discretion provide a relatively safer way to be involved. And although the Internet may expedite and ease interactions, its role should not necessarily be viewed as an alternative to physical, offline interactions. Gill et al. (2017) found that cases in which extremist activity was solely conducted online remain rare. The research noted that the role of the Internet was largely for instrumental purposes and that "there is little evidence to suggest that the Internet was the sole explanation prompting actors to decide to engage in a violent act" (p. 16). Koehler's work (2014) exploring the role of the Internet in radicalisation processes based on interviews with former right-wing extremists found that the Internet offered a venue that facilitated information exchange, ideological development and training opportunities, and shaped individuals' radicalisation pathways. Some interviewees cited the role of the Internet in facilitating the purchase of right-wing merchandise, music and clothing. Others, however, felt that, while online interactions were important, they were incomparable to attending rallies or having hard-copies of literature (Koehler 2014, pp. 122–123). The Internet may make some extremist activities more accessible and lower the threshold for involvement. However, it does not necessarily replace the significance or quality of offline engagements.

The Internet has also been instrumentalised for operational purposes, such as training, financing and attack planning. Furthermore, concerns have been raised about such operational activities shifting from the relatively more accessible and traceable surface web to the dark web, where governments, law enforcement agencies and tech firms have limited capacity to detect and disrupt activity (Malik 2018).

This highlights the need to look beyond the role of the Internet in violent extremist activity purely through the lens of indoctrination and ideology. Instead, there's a need to consider how the medium can facilitate operational and logistical activities related to violent extremism. On the relationship between the online and offline realms, Stevens and Neumann posit that it is more accurate to view the interaction as cyclical and complementary in nature: "Where radicalisation has a virtual component, that element needs to be seen as part of an iterative process through which events and developments in the real world are fed into cyberspace and vice versa" (2010, p. 13).

In the context of tackling online extremism, the Home Affairs Select Committee's report on radicalisation described the use of the Internet to promote radicalisation and terrorism as "one of the greatest threats that countries including the UK face" and that "the Internet has a huge impact in contributing to individuals turning to extremism, hatred and murder" (Home Affairs Select Committee 2016). Online radicalisation guidance published by the UK's National Counter Terrorism Security Office notes that "the Internet has transformed the way that terrorist organisations can influence and radicalise people" and that tackling extremist content on the Internet is "vital in countering the terrorist narrative and stopping offences that incite terrorism".[6] At the launch of the UK's updated Counter-Terrorism Strategy in 2018, former Home Secretary Sajid Javid highlighted the importance of eliminating safe spaces for terrorist propaganda online to prevent people from being "radicalised in a matter of weeks".[7] Similarly, UK Counter-Terrorism Policing Lead, Assistant Commissioner Neil Basu, has spoken about the availability of terrorist propaganda online for individuals "seeking to radicalise themselves and others", and that platforms must take greater

6 www.gov.uk/government/publications/online-radicalisation/online-radicalisation

7 www.bbc.com/news/uk-politics-44351841

responsibility for the threat posed by materials they host.[8] These views show that, at least in the UK context, the opinion among policymakers and law enforcement alike is that extremist and terrorist content online have a causal relationship with violence.

By contrast, researchers and academics are less convinced about the emphasis on online radicalisation, and subsequent counter-measures. Instead, they contend that responses should look beyond online-only measures, such as takedowns and counter-narratives, and prioritise a more holistic approach reflecting offline activity and interactions (Hamid 2018). In their work analysing online behaviours of convicted UK terrorists, Gill et al. concluded that "radicalisation should therefore be framed as cyber-enabled rather than cyber-dependent" (2015, p. 37).

Others consider the prevalence of ideologically extremist material online as subordinate to the attraction of participation in direct violence, access to weapons, and the twin pursuit of excitement and rejection of the mundane. Crone argues, "Young people who currently travel to conflict zones abroad are not necessarily illuminated by a radical religious ideology inciting them to engage in a foreign conflict" (2016, p. 594). She suggests accessing extremist ideological content online comes after the initial attraction, providing justification for an established desire to participate in violent extremism, rather than a driving force behind an individual's activity.

A RAND Europe study published in 2013 entitled 'Radicalisation in the Digital Era', which was granted access to several case files of individuals convicted on terrorism charges in the UK, was able to explore each individual's online activity prior to conviction or engagement with terrorist activity. The conclusions from the study were that "the Internet may enhance opportunities to become radicalised" but the study did not necessarily support the idea that the Internet accelerates radicalisation. Most significantly, it did not encounter any evidence to indicate that "the Internet is replacing the need for

8 www.counterterrorism.police.uk/neil-basu-welcomes-online-safety-measures/

individuals to meet in person during their radicalisation process" (Von Behr et al. 2013, p. xii). Instead, the findings suggest online activity complemented offline interactions.

Holbrook's study (2017) into the type of religious, political and ideological content unearthed during terrorism investigations in the UK focuses on the media environments of convicted terrorists. This approach offers insight into the influences and frames of reference that may have shaped individuals' involvement in terrorism. Holbrook's work highlights the diversity of content types and sources found in the possession of convicted terrorists. These include legal content and materials by popular clerics that could simply be described as religiously conservative rather than extremist.

The research landscape shows that, while there is an acknowledgement in academic circles that the Internet can play a role in extremist and terrorist behaviour offline, it must be viewed more as an enabler rather than a driver. The convenience and efficiency of the Internet facilitates group interactions and indoctrination processes, playing a contributory role in an individual's radicalisation. However, this engagement cannot be separated from offline communications, meetings and activities. As policymakers and law enforcement lean on causal explanations of the relationship between the Internet, radicalisation and violent extremism, more nuanced understandings of the role of extreme digital speech are obscured.

CASE STUDIES: VIOLENT EXTREMISTS AND INTERNET USE

This section provides a brief overview of cases in the UK, specifically Islamist extremists Khalid Masood, perpetrator of the 2017 Westminster Bridge attack; Ahmed Hassan, the teenager behind the bombing on a London Underground train at Parsons Green station; right-wing extremist Thomas Mair, who murdered Jo Cox MP; and Darren Osborne, the driver behind the attack on Muslim worshippers in Finsbury Park. These short case studies show the specific and often unique ways in which extreme digital speech and the Internet are implicated in political violence.

The report commissioned by the UK government into the 2017 terrorist attacks found that the Westminster Bridge attacker, Khalid Masood, used the Internet to conduct reconnaissance ahead of his attack. Searches of his devices revealed strong conservative religious inclinations, though his digital media collection was described as devoid of 'standard jihadi content' (Anderson 2017, p.14). Despite ISIS claiming responsibility for the attack, there was no evidence of Masood accessing or possessing its propaganda materials. Moreover, Masood was on the radar of British intelligence services since 2004 until his case was closed by MI5 in 2012 (Kirk 2018). Masood also had a history of violence and had spent time in prison, which is when he is believed to have converted to Islam and been radicalised. While there was evidence of extreme digital content in Masood's possession, it is inaccurate and imprudent to single this out in the context of other contributory factors that may have played a role in influencing his actions.

The Parsons Green bomber, Ahmed Hassan, an Iraqi asylum seeker, used ingredients purchased on Amazon and a bomb-making guide he found elsewhere to create an explosive device. Hassan listened to jihadi nasheeds[9] produced by ISIS that encourage violence. A college mentor referred Hassan to the 'Prevent' anti-terror programme after seeing messages on his phone (Dearden 2018). Hassan also previously told Home Office officials about being 'trained to kill' by ISIS while in Iraq. In Hassan's case, the significance of online activity in the ideological or indoctrination phase appears to be inconsequential, given his physical interaction with ISIS, whom he credited with training him. And his use of Amazon is an example of why research into terrorist exploitation of the Internet must look beyond ideological content and consider a more comprehensive range of threats.

Based on his Internet activity, Thomas Mair, who murdered Jo Cox MP in 2016, was found to have been fascinated by Nazism and white supremacist groups such as the Ku Klux Klan. Also, hard copies

9 Nasheeds are a form of religious music that involve rhythmic chanting and are often sung a capella or accompanied with percussion.

of extremist material were found at his home. During his trial, it was noted that Mair used the Internet to conduct research into political assassinations, including that of US President John F. Kennedy and Conservative MP Ian Gow (Cobain, Parveen and Taylor 2016; Casciani and Simone 2016). The extent to which Mair was radicalised online is more accurately understood when viewed alongside his offline activities. In letters sent to the editor of a South African pro-Apartheid publication as far back as 1988, Mair expressed his support for white supremacism and spoke of an impending violent struggle (Burgis 2016). Similarly, records show that Mair was a dedicated supporter of the National Alliance, a US neo-Nazi organisation, and he purchased manuals from the group containing instructions on how to build a pistol and explosive devices (Southern Poverty Law Center 2016). The evidence suggests Mair's extremist views were well-established *long before* the proliferation of the Internet and social media.

Finsbury Park attacker Darren Osborne, who drove his van into a crowd of Muslim worshippers outside an Islamic centre in north London, was also reported to have been radicalised online in just a matter of weeks after having engaged on social media with the views of far-right activists such as Tommy Robinson, former leader of the English Defence League, and Jayda Fransen, former deputy leader of Britain First (Dodd 2018; Rawlinson 2018). However, details about how Osborne's extremist views developed indicate that far from being solely motivated by content viewed online, a more traditional medium may have contributed towards his radicalisation. The 2017 BBC drama Three Girls, based on the true stories of victims of the Rochdale grooming scandal, had a major impact on Osborne. His former partner described him as being obsessed by Muslims after watching the show, holding all Muslims responsible for child sexual abuse and terrorist attacks. While it may be difficult to establish the precise sources of inspiration for Osborne's views, it is grossly inaccurate to assert that the Internet was the sole driver. Osborne's obsession with and contempt of Muslims after watching a BBC television programme demonstrates that individuals' views can just as plausibly be influenced by mainstream content, not just extreme digital speech.

While the Internet played *a role* in each of the case studies examined, the nature and extent of Internet activity differs in each instance. It is difficult, if not impossible, to suggest that the sole motivation, inspiration or driving force behind their violence was due to engagement with extreme digital speech online. The highly individualised patterns of engagement and nature of influence highlights the challenges governments and tech companies face in responding to extremist and terrorist use of the Internet, where the impact of particular content may enable or inspire some but not others, and where it is almost impossible to implement policies or practices that are tailored to individual experiences.

MOVING FORWARD

How extremist content online influences or impacts individuals and the efficacy of responses to extremist and terrorist content online requires greater research focus (see Schmid and Forest 2008). More empirical assessment and evaluation are needed around online engagement and linkages to violent extremist activity. Governments, multilateral organisations, tech platforms, law enforcement agencies and civil society organisations alike need to work collaboratively to increase the precision and efficacy of responses. To better understand the relationship between violent extremist content online and activity offline, researchers could shift from exploring *what* and *where* and instead pursue answers to *how* and *why*. A sizable evidence base has emerged in recent years for understanding what type of content extremist groups are producing (videos, magazines, lectures and so on) and which platforms (social media, encrypted messaging, search engines, the dark web and so on) are being used to circulate materials. However, there is a dearth of literature looking at pathways that link online and offline behaviours to violent action. One major way in which the methodological challenges in understanding user interaction with extreme digital speech could begin to be addressed is with greater access to case files of individuals convicted of terrorism offences, in particular their Internet history and online activity.

Governments should seek to co-operate more closely with the research community through providing access to the Internet histories of convicted violent extremists, as this could significantly help bridge the gap between understanding online and offline behaviours. Evidence relating to an individual's online activity should be shared with a wide pool of vetted research partners to help establish a substantial, representative evidence base that maps the interplay between online and offline activity, as well as highlighting the spectrum of online activity over time. The RAND study cited above (Von Behr et al. 2013) benefited from access provided by the UK Home Office and its Office for Security and Counter-Terrorism (OSCT). Improved access to terrorism case files, with the necessary safeguards and privacy mechanisms in place, will help improve the understanding of how violent extremist perpetrators may, or may not, be influenced by online content, helping to piece together a more comprehensive picture of the relationship between Internet activity and extremist violence.

Consensus on the precise role of the Internet and violent extremist content online in individual radicalisation may remain elusive and continue to divide opinion, but the role of digital content cannot be ignored. As illustrated by the case studies of violent extremist perpetrators, not only are individuals subject to influence from different sources – online and offline – the nature of online engagement varies from case to case. Furthermore, the role of personal circumstances, mental health issues and perceived grievances cannot be downplayed as factors in an individual's radicalisation. In this milieu of personal circumstances and online influences, the overwhelming focus of counter-efforts has been on the latter, whether through content removal or through counter-narrative campaigns. There is, however, little evidence to suggest that these responses are producing any substantial changes in behaviour or thought. Singling out the Internet as the source or driver of extremist violence is dishonest and disingenuous, and it risks creating the expectation that addressing matters online will inevitably resolve matters offline.

References

Anderson, D. 2017. *Attacks in London and Manchester* (Rep.). Home Office.

Burgis, T. 2016. 'Thomas Mair: The Making of a Neo-Nazi Killer'. *The Financial Times*, 23 November. Retrieved from: www.ft.com/mair

Casciani, D. and Simone, D. 2016. 'Thomas Mair: Extremist Loner Who Targeted Jo Cox'. *BBC News*, 23 November. Retrieved from: www.bbc.co.uk/news/uk-38071894

Cobain, I., Parveen, N. and Taylor, M. 2016. 'The Slow-Burning Hatred that Led Thomas Mair to Murder Jo Cox'. *The Guardian*, 23 November. Retrieved from: www.theguardian.com/uk-news/2016/nov/23/thomas-mair-slow-burning-hatred-led-to-jo-cox-murder

Counter Terrorism Policing. 2019. *Neil Basu Welcomes Online Safety Measures*, April. Retrieved from: www.counterterrorism.police.uk/neil-basu-welcomes-online-safety-measures/

Crone, M. 2016. 'Radicalization Revisited: Violence, Politics and the Skills of the Body'. *International Affairs*, 92 (3), pp. 587–604.

Dearden, L. 2018. 'Parsons Green Bomber's Isis Inspiration Was Missed by Police Investigation'. *Independent*, 21 March. Retrieved from: www.independent.co.uk/news/uk/crime/parsons-green-attack-police-missed-isis-evidence-bombing-tube-a8267601.html

Dodd, V. 2018. 'How London Mosque Attacker Became a Terrorist in Three Weeks'. *The Guardian*, 1 February. Retrieved from: www.theguardian.com/uk-news/2018/feb/01/finsbury-park-london-mosque-van-attack-darren-osborne-makram-ali

European Commission. 2018. *A Europe that Protects: Commission Reinforces EU Response to Illegal Content Online* [press release]. Retrieved from: www.europa.eu/rapid/press-release_IP-18-1169_en.htm

Gill, P., Conway, M., Corner, E. and Thornton, A. 2015. 'What Are the Roles of the Internet in Terrorism?' (Rep.). *VOX-Pol Network of Excellence*.

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), pp. 99–117.

Hamid, N. 2018. 'Don't Just Counter-Message; Counter-Engage'. *International Centre for Counter-Terrorism*, 28 November. Retrieved from: www.icct.nl/publication/dont-just-counter-message-counter-engage/

Holbrook, D. 2017. *What Types of Media Do Terrorists Collect?* (Rep.). International Centre for Counter-Terrorism.

Home Affairs Select Committee. (2016). Radicalisation: The counter-narrative and identifying the tipping point (Rep.). House of Commons, Home Affairs Select Committee. Retrieved from: www.publications.parliament.uk/pa/cm201617/cmselect/cmhaff/135/135.pdf

Home Office (2018). 'Home Secretary Announces New Counter-Terrorism Strategy'. Retrieved from: www.gov.uk/government/speeches/home-secretary-announces-new-counter-terrorism-strategy

Kirk, T. 2018. 'MI5 Saw Westminster Attack Khalid Masood As Threat in 2010, Agent Tells Inquest'. *Evening Standard*, 26 September. Retrieved from: www.standard.co.uk/news/crime/mi5-saw-westminster-attacker-khalid-masood-as-threat-in-2010-agent-tells-inquest-a3946536.html

Koehler, D. 2014. 'The Radical Online: Individual Radicalization Processes and the Role of the Internet'. *Journal for Deradicalization*, Winter 2014/15 (1), pp. 116–134.

Malik, N. 2018. *Terror in the Dark* (Rep.). The Henry Jackson Society.

Meleagrou-Hitchens, A. and Kaderbhai, N. (2017). Research Perspectives on Online Radicalisation (Rep.). VOX-Pol.

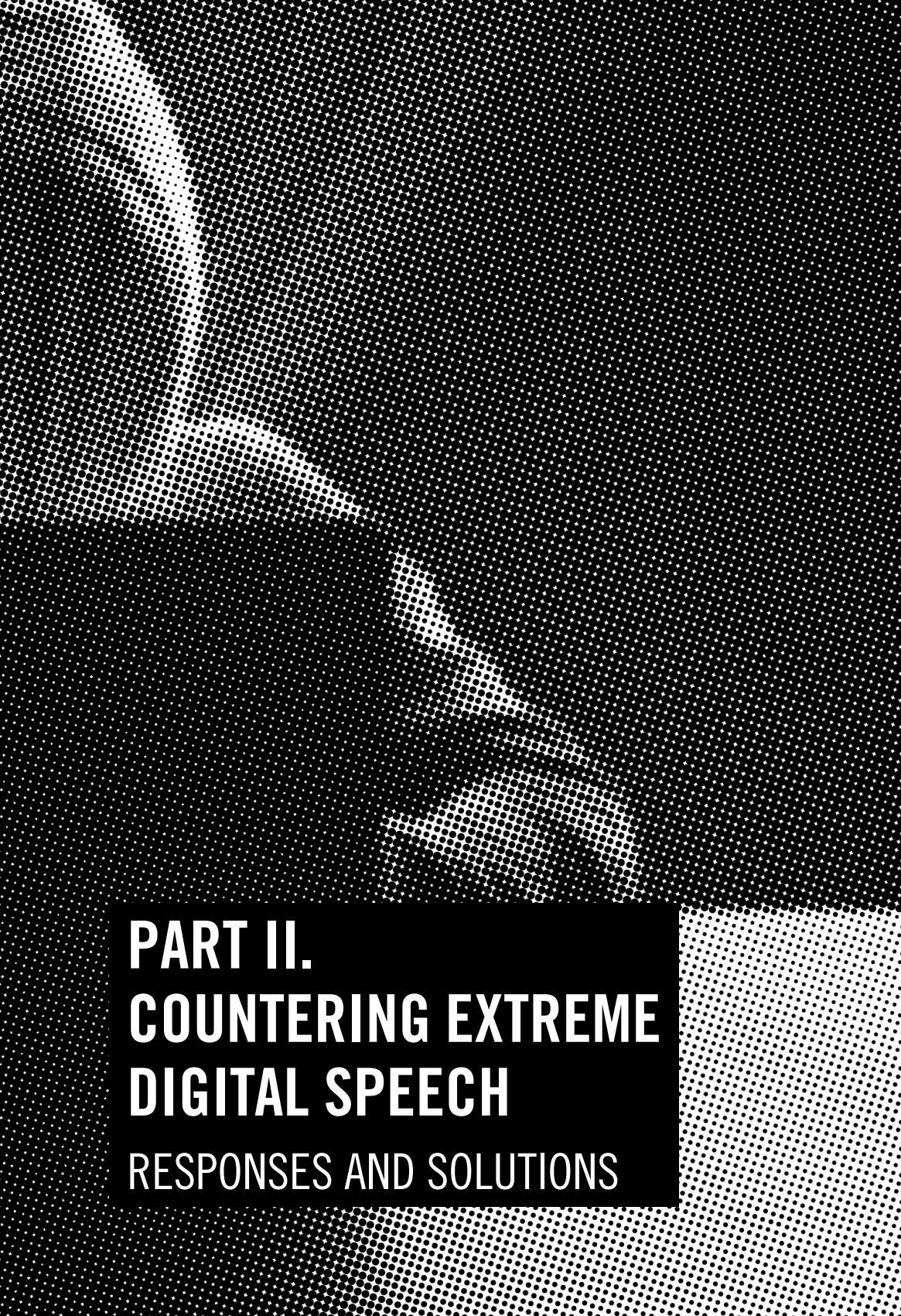
Rawlinson, K. 2018. 'Finsbury Park-Accused Trawled Far-Right Groups Online, Court Told'. The Guardian, 23 January. Retrieved from: www.theguardian.com/uk-news/2018/jan/23/finsbury-park-accused-wanted-to-kill-all-muslims-court-told

Schmid, A.P. and Forest, J.J. 2018. 'Research Desiderata: 150 Un- and Under-Researched Topics and Themes in the Field of (Counter-)Terrorism Studies – a New List'. Perspectives on Terrorism, 12 (4), pp. 68–776.

Southern Poverty Law Center. 2016. Alleged Killer of British MP Was a Longtime Supporter of the Neo-Nazi National Alliance. 16 June. Retrieved from: www.splcenter.org/hatewatch/2016/06/16/alleged-killer-british-mp-was-longtime-supporter-neo-nazi-national-alliance

Stevens, T. and Neumann, P. 2010. Countering Online Radicalisation (Rep.). International Centre for the Study of Radicalisation.

Von Behr, I., Reding, A., Edwards, C. and Gribbon, L. 2013. Radicalisation in the Digital Era (Rep.). RAND Corporation.



PART II. COUNTERING EXTREME DIGITAL SPEECH

RESPONSES AND SOLUTIONS

AUTOMATED DETECTION OF TERRORIST AND EXTREMIST CONTENT

John Gallacher

Given the huge volume of content posted online and to social media platforms, it is unfeasible to have human moderators manually inspect every single item. In order to detect terrorist or extremist content, one of the most prominent approaches is to use automated tools to scan all online content and automatically detect content that reaches certain thresholds and to remove it from the platforms. This detection is increasingly being done through the use of 'intelligent' tools based on machine learning. After discussing how automatic detection using machine learning works at a system level, this essay explores how these tools perform when deployed in the real world and the existing limitations and challenges that they face.

HOW DOES AUTOMATIC DETECTION WORK?

Machine learning is a branch of artificial intelligence (AI) that aims to emulate human decision-making ability in computers. By looking for patterns in datasets and using statistical programming to 'learn' over time, computers can update the criteria used to make decisions based on growing levels of available information. One branch of machine learning that is gaining popularity is known as 'deep learning'. This uses computer simulations of biological neurons to create 'artificial neural networks' (ANNs). In these networks, one neuron connects to a number of others, and information is transferred through the network by patterns of activation in connected neurons. By adjusting the weights of the connections between neurons using feedback from trial and error, artificial neural networks can extract meaningful patterns from datasets (Schmidhuber 2015). With enough training-set data to learn from, these neural networks can categorise images, text, speech, videos and so on into distinct classes. These

methods could help differentiate types of animals in photos, for example, or potentially to distinguish extreme digital speech from mainstream communication online (Johnson et al. 2017).

Given the vast amount of data that has become available since the explosion of the Internet, machine-learning algorithms have more data than ever to learn from and perform ever-more advanced functions aiming to emulate human behaviour – and sometimes surpass it in terms of speed and accuracy (Wang et al. 2015). This availability of data makes these methods highly valuable tools for blocking extremist content, as they can be scaled in concert with the rising levels of extremist material, with potential to relieve growing pressure on human moderators (Bickert and Fishman 2017).

CURRENT LEADING EFFORTS

Due to this opportunity, many social media and technology firms have begun to employ machine-learning algorithms on their platforms to automatically block and delete content that is classified as extremist. When extremist material is blocked on one platform, there is a concern that terrorists will simply move to another. In order to prevent this situation, technology firms announced a collaboration in 2016. Initially set up by Google, Microsoft, Facebook and Twitter, the primary focus of the Global Internet Forum to Counter Terrorism (GIFCT) is to create a shared database of extremist content to block repeated uploads of the same material (Google Blog 2017). This is done by creating a cryptographic hash or 'digital fingerprint' of each identified image or video clip, and automatically blocking database matches that are found when content is uploaded to any platform. So far, the database contains 88,000 digital fingerprints, and smaller platforms – Ask.fm, Cloudinary, Instagram, Justpaste.it, LinkedIn, Oath and Snapchat – have been added to the data sharing group (Lee 2018).

An alternative, though similar, approach was announced in February 2018 by the UK government by revealing a partnership with artificial intelligence firm ASI Data Science, now known as Faculty (ASI Data Science Blog 2018). ASI Data Science created a tool aiming

to accurately detect online jihadist video content, particularly ISIS propaganda. The tool employs machine learning to analyse a range of subtle patterns within videos and determine whether these patterns match those found in ISIS propaganda. Importantly, it can distinguish original content from videos that discuss the same imagery in a different context, such as news reports or commentary over previous footage. The reported accuracy of this tool is 94% with a false positive rate of 0.005%. ASI Data Science claims that, by employing this method, one trained analyst would be able to monitor all content uploaded to a platform the size of the video-sharing platform YouTube. Without such a tool, it would take approximately 20,000 people to perform the same task. The tool is not intended as the only level of moderation, but would, rather, flag videos for human consideration.

Not all extremist content is as explicit as this, however, and when it comes to the task of detecting content that falls under the definition of extreme digital speech, a different approach is needed. Language in the category of extreme digital speech may cover a broader range of issues, including hateful or dehumanising speech without explicit calls to violence. In late 2017, Google launched 'Perspective API', a machine-learning tool that uses natural language processing to score the perceived impact that a comment might have on an online conversation. Within this model, comments that are ruder, more disrespectful or more aggressive receive a higher 'toxicity' score (Wulczyn, Thain and Dixon 2016). The tool was developed through the manual coding of millions of comments from different online platforms on a scale from 'very toxic' to 'very healthy', and this training set was used to teach the tool to classify comments in new and unseen conversations. Since being launched in 2017, the tool has been used by Wikipedia to study the effect of personal attacks on editors, and by The New York Times to automatically highlight abusive comments within online discussions. The stated aim of the tool is to reduce the burden on moderators by removing the need to manually review every comment, instead allowing them to focus only on comments with high toxicity scores.

AUTOMATIC DETECTION SUCCESS

Automatic content removal has been highly successful in detecting and removing certain types of content. Facebook reports that it automatically removes 99% of ISIS and al-Qaeda material uploaded to the platform (Bickert and Fishman 2018). Similarly, YouTube reports that 98% of the videos that are removed for violating violent extremism rules are detected by an automated system (YouTube Official Blog 2017). During the second six months of 2017, Twitter removed 275,000 accounts for violations related to promotion of terrorism, and 93% of these accounts were flagged for removal by automated tools (Twitter Transparency Report 2018).

This undoubtedly represents a success for the major social media platforms (Macdonald 2018). Automatic content detection tools are allowing more content to be removed, and for this to be done more quickly and with greater coverage. As a result, these platforms have become a much more difficult place for terrorist organisations to operate (Conway et al. 2017).

EXISTING CHALLENGES FOR AUTOMATIC DETECTION

Despite this initial success, four primary problems with automatic content deletion still exist:

1. How to deal with false positives.
2. How to prevent malicious attacks on the tools. This is especially difficult as efforts to make these tools open source and transparent increase opportunities for exploitation by revealing potential vulnerabilities and weaknesses in the source code.
3. How to apply these tools when groups move to more encrypted spaces and smaller social media platforms.
4. How to appropriately apply tools built with a focus on one domain of extremism to other forms of extremism where the use of online spaces is markedly different.

How to Deal with False Positives

The claimed accuracy of the tool created by ASI Data Science is 94% with a false positive rate of 0.005%. However, even these very promising numbers do present a problem when expanded to cover the sheer volume of data available online. Consider the following: In a situation where 100,000 videos are uploaded to a platform, 100 of which are of extremist content, the ASI Data Science tool would detect ninety-four of the extremist videos correctly and fail to detect six. Additionally, the machine learning classifier would wrongly identify five regular videos as extremist. For relatively small samples such as this, these numbers seem acceptable. However, as the proportion of extremist content to regular content decreases, the number of false positives becomes problematic. If the classifier is presented this time with 10 million videos, and again 100 of these are extremist content, then, as before, ninety-four extremist videos will be accurately detected and six missed. However, the number of false positives will increase to 500. In this situation, over five times as many videos are incorrectly flagged as extremist propaganda than are correctly identified. Given that over 400 hours of video are uploaded to YouTube every minute, this false positive rate could lead to a large number of false positives in the real world, and cause either increased strain on human moderators or the removal of genuine content.

Additionally, even in cases where a video is successfully classified as containing extremist material, issues can still arise given that automated systems cannot take into account a wider context. In 2017, YouTube removed thousands of videos documenting the conflict in Syria as these videos triggered automatic classifiers (Browne 2017). Human rights advocates claim these videos contained vital evidence of human rights abuses and therefore the removal jeopardises the chances of criminal proceedings being successful.

How to Prevent Malicious Attacks on the Tools

Automatic content classifiers are susceptible to malicious 'attacks' on the systems. Ilyas et al. (2018) demonstrated how Google's state-of-the-art InceptionV3 image-classification tool is vulnerable

to being tricked into miscategorising 'adversarial example images'. These images are carefully altered examples that cause misclassification through the addition of computer-generated noise in the image that is often invisible to the human eye, but is effective in tricking the machine. In one example, a model of a turtle is 'perturbed' so as to be classified as a rifle from any angle. Such attacks on automatic content-removal tools could prove damaging, either through the manipulation of images to bypass the moderators or through artificial over-classification that swamps the system. The proliferation of technology has reduced the barriers to carrying out this type of activity, and the computer code required to perform this image manipulation is available publicly on GitHub (Ilyas et al. 2018).

Similarly, Hosseini et al. (2017) have successfully 'deceived' the Google Perspective project through minor changes to spelling and grammar in the comments that the tool is asked to classify. These changes lead to a lower toxicity judgement by the machine, but no decrease in the impact on a human recipient of the message. These examples demonstrate how automatic classifiers require constant updating and expansion in the training set in order to keep pace with a constantly adapting digital landscape.

These issues are not restricted to academic discussions, and there are a growing number of real-world cases where simple changes to images or videos have been shown to bypass fingerprinting techniques by changing the content just enough so that machines won't recognise it as identical. This was demonstrated at scale in the wake of the March 2019 shooting at the Al Noor mosque in Christchurch, New Zealand. The shooter 'livestreamed' the attack on Facebook (Sonderby 2019), and whilst the original video was quickly removed, in the following twenty-four hours Facebook logged 1.5 million attempts to re-upload the same content. Using automated detection tools, Facebook blocked 1.2 million of these at the point of upload, as the content matched a fingerprint of the original which they had also shared with the GIFCT. However, the remaining 300,000 versions of the video were not immediately detected. The reasons for these false negatives are reported to be that users made small modifications to the videos, including changing the size and length of the clips,

adding logos or watermarks, or even turning the video into an animation (Dwoskin and Timberg 2019). These changes are very easy for users to make using simple video editing software and make the success of the tools drop dramatically. In the time it takes to detect them with other methods, such as users flagging the content manually, the content can be exposed to a wide audience.

In a similar way, digital text is increasingly embedded within images that are shared online. Sometimes this is in order to directly bypass text censorship, but more often it is part of the phenomena of sharing memes – comical captions on photos or cartoons that are often intended to caricature elements of human behaviour. Whilst most memes are shared in a positive way for benign purposes, others have gained negative or hate-filled connotations. This latter category of memes often originates within fringe online communities as a method of disseminating hate speech (Zannettou et al. 2018). As a response to this challenge of hate speech embedded within image/text combinations, Facebook has built 'Rosetta'. This is a machine-learning system that extracts the text within an image file and inputs this into a recognition model that has been trained to understand the context of the text and the image together (Borisyuk et al. 2018). Facebook claims that, by analysing the image and text together, it can understand the text in the way it was intended by the author and, as such, proactively identify inappropriate or harmful content. This method is able to process more than one billion memes a day, but whether it can detect the nuance used by niche Internet subcultures remains to be seen.

Encrypted Channels and Automatic Content Detection

The third challenge to automatic content-removal tools is how to apply them as extremist groups move to operate on encrypted communication channels such as WhatsApp, Telegram or Signal. This creates two hurdles: one for the platform creators and one for third parties such as intelligence or law-enforcement agencies.

Due to the ways in which encryption scrambles and masks the content of messages, automatic content-detection tools are much harder, and sometimes impossible, to employ. For platforms, this means that they cannot easily use automatic detection tools to prevent extremist content from being uploaded, as the encrypted file will not match a digital fingerprint stored in a repository even if the content is identical. One solution may involve deploying the monitoring tools 'directly' on users' devices so that they can scan content before the encryption is applied. However, this would lead to challenges of privacy infringement and corporate surveillance, and would likely cause users to simply switch to an alternative platform.

For law-enforcement or intelligence agencies, the challenge of encrypted platforms is that of data access. The conversations on encrypted platforms are typically 'private' and available only to certain members, who are added to the encrypted protocol when invited to join. This means that mass detection techniques across a platform as a whole are not possible and much more effort is needed to infiltrate the specific 'chats' directly.

Applying the Tools to Other Forms of Extremism

Many of the automated detection tools discussed here have been built with a focus on jihadist content, and the machine-learning techniques tuned specifically to detect the imagery, language and visual cues associated with this type of extremism. As the global threat of extremism changes over time, how to repurpose these existing tools to detect other forms of extremist content poses a challenge.

This is especially challenging for the detection of right-wing extremist content. The types of language and visual content used by these groups are markedly different from jihadist groups, often with numerous layers of irony, 'in-jokes' or obscure references which repurpose mainstream ideas (Evans 2019). This combination makes online spaces used by right-wing extremist groups difficult for an outsider to understand, and even harder for a machine to correctly interpret without increasing the rate of false positives. This also

creates a risk where using automated tools unfit for the purpose will cause platforms to fall into traps laid for them by digital-savvy groups, causing misattribution or failure to spot signs of potential violence.

Additionally, these far-right groups have been shown to increase their use of coded language, euphemisms and punctuation in place of alphabetic characters, in order to evade detection from automatic classifiers (Magu and Lao 2018). As online groups move from fringe Internet platforms towards the mainstream, automatic classifiers will need to be trained to take into account all of this group history to try and make sense of a language used (Zannettou et al. 2018).

CONCLUSION

Whilst good progress has been made in developing automated technologies to detect and remove extremist content, these tools are still in their infancy and represent only the first step in a constantly changing landscape. It is vital that platforms, governments, investors and the public realise that automatic detection is not a problem that can be 'solved'. Rather, it is something that will require constant innovation and development to avoid falling behind changes in the way extremist groups use the Internet to spread extreme digital speech.

Future developments may be to include information from multimodal domains to improve the contextual understanding of the messages or to detect coded messages (Rudinac et al. 2017). Facebook has already moved to implement these changes in the wake of the New Zealand shooting, adding audio files as well as visual content to the shared cryptographic hash database in an attempt to better detect future re-uploads of the same content. More efforts in this area are likely to improve the success of these tools.

There are also a growing number of academic researchers working in this area using innovative applications of automated and semi-automated techniques to detect terrorist and extremist content online. This research makes use of novel features such an individual's specific writing style to better detect offensive speech in a real-world context (Chen et al. 2012) or to detect when codewords are used in place of overt hate words (Taylor et al. 2017). Alternatively,

the detection of extreme digital speech can be improved by identifying when the same individual is using multiple online aliases to spread extreme digital speech (Dahlin et al. 2012) or by including temporal aspects to measure the change in a user's propensity to extremism over time (Scrivens, Davies and Frank 2018). These applications are still in development, and more work will be required before they can be deployed in a real-world system. A closer relationship in the future between the platforms which hold the data and researchers is likely to lead to advances in extremist content detection whilst preserving privacy.

References

ASI Data Science. 2018. 'Flagging Extremist Content Online', February. Retrieved from: www.blog.asidatascience.com/flagging-extremist-content-online [Accessed 21 November, 2018].

Bickert, M. and Fishman, B. 2017. 'Hard Questions: How We Counter Terrorism', 15 June. Retrieved from: www.newsroom.fb.com/news/2017/06/how-we-counter-terrorism [Accessed 17 June, 2019].

Bickert, M. and Fishman, B. 2018. 'Hard Questions: How Effective Is Technology in Keeping Terrorists off Facebook?', 23 April. Retrieved from: www.newsroom.fb.com/news/2018/04/keeping-terrorists-off-facebook [Accessed 21 November, 2018].

Borisyuk, F., Gordo, A. and Sivakumar, V. 2018. 'Rosetta: Large Scale System for Text Detection and Recognition in Images'. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 71–79.

Browne, M. 2017. 'YouTube Removes Videos Showing Atrocities in Syria'. New York Times, Retrieved from: www.nytimes.com/2017/08/22/world/middleeast/syria-youtube-videos-isis.html [Accessed 11 July, 2018].

Chen, Y., Zhu, S., Zhou, Y. and Xu, H. 2012. 'Detecting Offensive Language in Social Media to Protect Adolescents'. 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, pp. 71–80.

Conway, M., Khawaja, M., Lakhani, S., Reffin, J., Robertson, A. and Weir, D. 2017. 'Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts'. *VOX-Pol Network of Excellence*.

Dahlin, J., Johansson, F., Kaati, L., Martenson, C. and Svenson, P. 2012. 'Combining Entity Matching Techniques for Detecting Extremist Behavior on Discussion Boards'. 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 850–857.

Dwoskin, E. and Timberg, C. 2019. 'Inside YouTube's Struggles to Shut Down Video of the New Zealand Shooting – and the Humans who Outsmarted its Systems'. *The Washington Post*, 18 March. Retrieved from: www.washingtonpost.com/technology/2019/03/18/inside-youtubes-struggles-shut-down-video-new-zealand-shooting-humans-who-outsmarted-its-systems [Accessed 17 June, 2019].

Evans, R. 2019. 'Shitposting, Inspirational Terrorism, and the Christchurch Mosque Massacre'. *Bellingcat*, 15 March. Retrieved from: www.bellingcat.com/news/rest-of-world/2019/03/15/shitposting-inspirational-terrorism-and-the-christchurch-mosque-massacre [Accessed 17 June, 2019].

Google Public Policy. 2017. 'Update on the Global Internet Forum to Counter Terrorism', 04 December. Retrieved from: www.blog.google/around-the-globe/google-europe/update-global-internet-forum-counter-terrorism [Accessed 21 November, 2018].

Hosseini, H., Kannan, S., Zhang, B. and Poovendran, R. 2017. 'Deceiving Google's Perspective API Built for Detecting Toxic Comments'. arXiv:1702.08138.

Ilyas, A., Engstrom, L., Athalye, A. and Lin, J. 2018. 'Black-box Adversarial Attacks with Limited Queries and Information'. arXiv:1804.08598v2.

Johnston, A.H. and Weiss, G.M. 2017. 'Identifying Sunni Extremist Propaganda with Deep Learning'. 2017 IEEE Symposium Series on Computational Intelligence, pp. 1–6.

Lee, D. 2018. 'Tech Firms Hail 'Progress' on Blocking Terror'. *BBC News*, 8 June. Available at: www.bbc.com/news/technology-44408463 [Accessed 11 July, 2018].

Macdonald, S. 2018. 'How Tech Companies are Successfully Disrupting Terrorist Social Media Activity'. The Conversation, 6 June. Retrieved from: www.theconversation.com/how-tech-companies-are-successfully-disrupting-terrorist-social-media-activity-98594 [Accessed 11 July, 2018].

Magu, R. and Luo, J. 2018. 'Determining Code Words in Euphemistic Hate Speech Using Word Embedding Networks', Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pp. 93–100.

Rudinac, S., Gornishka, I. and Worring, M. 2017. 'Multimodal Classification of Violent Online Political Extremism Content with Graph Convolutional Networks', ACM Thematic Workshops, pp. 245–252.

Schmidhuber, J., 2015. 'Deep Learning in Neural Networks: An Overview'. *Neural Networks* 61, pp.85–117.

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), pp. 39–59.

Sonderby, C. 2019. 'Update on New Zealand', 18 March. Retrieved from: www.newsroom.fb.com/news/2019/03/update-on-new-zealand [Accessed 17 June, 2019].

Taylor, J., Peignon, M. and Chen, Y. 2017. 'Surfacing Contextual Hate Speech Words within Social Media'. arXiv:1711.10093.

Twitter Transparency Report. 2018. 'Government TOS Reports'. Retrieved from: www.transparency.twitter.com/en/gov-tos-reports.html [Accessed 21 November, 2018].

Wang, H., Tian, F., Gao, B., Bian, J. and Liu, T.Y. 2015. 'Solving Verbal Comprehension Questions in IQ Test by Knowledge-Powered Word Embedding', arXiv:1505.07909.

Wulczyn, E., Thain, N. and Dixon, L. 2016. 'Ex Machina: Personal Attacks Seen at Scale' Proceedings of the 26th International Conference on World Wide Web, pp. 1,391–1,399. International World Wide Web Conferences Steering Committee.

YouTube Official Blog. 2017. 'Expanding Our Work against Abuse of Our Platform', 4 December. Retrieved from: www.youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html [Accessed 21 November, 2018].

Zannettou, S., Caulfield, T., Blackburn, J., De Cristofaro, E., Sirivianos, M., Stringhini, G. and Suarez-Tangil, G. 2018. 'On the Origins of Memes by Means of Fringe Web Communities'. *Proceedings of the Internet Measurement Conference*, pp. 188–202.

HUMAN ASSESSMENT AND CROWDSOURCED FLAGGING

Zoey Reeve

Artificial intelligence (AI) is undoubtedly useful in the battle against online terrorist material. However, it is a relatively blunt tool in some regards, particularly where the message presented is buried in complex, historical and/or religious doctrine, or does not include graphic imagery or symbols representing a proscribed terrorist organisation. The UK government and EU have placed increasing pressure on Internet providers and platforms to do more to tackle online terrorist material, bolstering this objective with the establishment of policing units such as the UK Metropolitan Police, Counter-Terrorism Internet Referral Unit (CTIRU), and the European Internet Referral Unit (EU IRU). Both units are small, with CTIRU consisting of about 10–15 officers (personal correspondence 2018), and EU IRU consisting of about 20 officers (Europol 2016b). Case officers are specialist staff seeking out and facilitating the removal of terrorist material from the online space. These IRUs work in co-operation with Internet platforms and providers, who may voluntarily remove illegal content. Platforms[10] such as Google, YouTube, Twitter and Facebook also have their own teams of human assessors, who are in communication with units such as CTIRU and EU IRU. These human assessors work to identify, assess and remove terrorist material from the online space. However, an important feature of the identification of such material comes from the public, who flag this material to the relevant organisation. This section of the report will focus on the human-assessment

10 This section of the report deals primarily with these large platforms or providers. However, it is important to recognise that there are a huge number of other platforms or providers that will be impacted by this issue. Smaller platforms or providers are unlikely to have the resources to deal with the problem in the way that these larger organisations do.

aspect of content moderation, particularly the policies and guidelines that are used by units such as CTIRU and platforms such as Twitter or YouTube to determine how a piece of material is dealt with.

WHAT MATERIAL MAY BE REMOVED

There is a general consensus across the major platforms that terrorist material should not be hosted. There are a number of detailed and largely similar policies tied up in the terms and conditions of these sites, which provide sites with the power to prevent such material from being hosted, or to remove it. The policies most relevant to this report are those that deal with terrorism and hate speech. Platforms such as Facebook readily admit that these policies are constantly re-evaluated (Bickert 2018). A talk given by Monika Bickert (2018), who is a policy lead for Facebook, revealed that the platform has around sixty experts who craft these policies, and sub-teams of policy experts within that team, one of which is dedicated to counter-terrorism and is made up of experts and practitioners in the field. Bickert states that Facebook recognises that, although some material is clear cut and easily identifiable as terrorist, and therefore removable via Facebook policies, most cases are not clear cut and require a careful consideration of potential harms that could occur in the offline world. In general, for platforms and providers, violence, or threats to violence, are to be prohibited. This includes wishing harm, death or disease on others, and it is here that threatening or promoting terrorism is identified as an issue (Facebook 2018c; Google 2018a; Twitter 2018b; YouTube 2018e). Essentially, users must not support or affiliate with organisations that engage with, promote or glorify terrorism. Terrorism, however, is distinct from extremism, the latter which is inherently relational and rooted on an ever-shifting centre – and is also not illegal.

For law enforcement units, it is legislation that forms the guidelines underpinning the assessment of the material and its removal – providing that the platform hosting the material falls under the legal jurisdiction. In the British case, for instance, material is assessed by individual CTIRU team members in accordance with the Terrorism

Act 2006. Online material that breaches the Terrorism Act 2006 typically falls into three overlapping categories that may be considered as constituting potential harm: material that glorifies and/or incites terrorism, material that provides instruction on how to commit acts of terrorism, and/or material that is produced by proscribed groups. CTIRU case officers primarily use the Terrorism Act to assess the material that they either find or to which they are referred, although they may also refer to other legislation or pass different types of offensive material to other relevant units. A similar process is used by other IRUs. Material that is extremist may not breach the Terrorism Act 2006 and may either be left, reported or reported to another team that deals with hate crime.

CONSEQUENCES FOR BREACHING POLICIES

The most obvious consequence of breaching a policy is that the material will be removed from the relevant platform. However, this is not the only consequence. Facebook has three ways of dealing with material that is flagged as breaching its policies: it can ignore the flag (decide that it is not a breach); it can delete the material (because it has been assessed as a breach); or it can mark the material as disturbing (Mark As Disturbing: MAD). This latter category is flagged when the material seems to occupy the grey area between breaching and not breaching a policy and includes restricting the content and presenting a warning tab to those wishing to view the content. MAD is used typically for material that is presented for the purposes of 'raising awareness' or to express extreme views, and is perhaps the most contestable decision (Dispatches 2018a). For instance, Facebook often MADs videos that have been created and uploaded for the purpose of raising awareness of child abuse, which has a number of ethical and legal implications which are outside of the scope of this essay. Similarly, YouTube issues 'strikes' in response to policy violations (YouTube 2018d). Videos or posts that violate policies can be removed, with explanation, and depending on the severity of the offence, the user's access to certain features and the capacity to post new content may be inhibited. Ultimately, a user's account can be

terminated, and their capacity to access, create or maintain any other YouTube channel will be impeded (YouTube 2018a).

Google takes action in a number of ways for policy violations (Google 2018a). This may include restricting access to the problematic content or removing it entirely. For Twitter, problematic content can be deleted. If users themselves do not delete problematic content, then said users will be unable to make new posts or interact with other users. Twitter can limit a user's capacity for creating new posts and interacting with other users temporarily in response to a violation, or the account can be suspended permanently. Importantly, all platforms make an exception for material that might be educational, documentary or artistic in nature. Thus, the context in which the material is presented influences how it is assessed and whether it is removed. Material that may be deemed 'removable' in some contexts is allowed to remain in place as if it is somehow less potentially harmful than when placed in an alternative context (in other words, uploaded by a user who supports a terrorist group, or non-terrorist material uploaded by a terrorist group). These intermediate strategies may encourage users to think more carefully about the material that they are uploading, educating them as to the policies and conditions held by the platform and why they are in place. However, these restrictions – whilst not as silencing as terminating the user's account completely – may be seen as unjust and punitive. This could potentially result in such users seeking less strictly moderated platforms, which may therefore expose them to more of this material.

HOW IS MATERIAL FOUND?

The larger Internet platforms and providers typically find material with AI, although this material has tended to largely be Islamist extremist in nature. According to Facebook, 99% of al-Qaeda and ISIS material uploaded was removed prior to it ever being found by users (Bickert 2018). Internet platforms and IRUs are not so successful in removing far-right extremist material, however. Nevertheless, human assessors continue to have an important role. Twitter, YouTube, Google and Facebook, for instance, all have human

moderators and user-focused tools. Looking at these platforms' websites, it is not always clear how these tools are used in the process of human assessment. A recent Dispatches investigation by Channel 4 has shed light on the process by which Facebook's frontline human moderation team assess material, and what they do with it (Dispatches 2018a). This process involves users identifying problematic pages with an explanation for why the complaint was launched using the tool for review, before this complaint becomes a ticket in a queue which is to be assessed within 24 hours. This type of approach is similarly pursued by the other major platforms. Material is typically identified upon being reported by users or members of the public. Indeed, platforms such as Google, YouTube and Facebook are unequivocal in their need for users to raise the alarm to inappropriate content (Facebook 2018a; Google 2018a; YouTube 2018c).

Tools for users to alert the platforms to potentially problematic material tend to be similar. Flagging a video on YouTube, for instance, requires users to sign in to YouTube, click on the 'More' tab beneath the problematic video, and choose the 'Report' option. This 'Report' will request information about the video, including the viewer's rationale for flagging the material (YouTube 2018c). Flagging problematic Twitter profiles is a very similar process (Twitter 2018a), as it is on Facebook (Facebook 2018b) and on Google (Google 2018b). For these platforms, then, human flagging of harmful material occurs largely via crowdsourcing.

For units like CTIRU, crowdsourced flagging makes up only one source of material. Material is found via three methods. One, officers may actively search for material using various tools, including web crawlers (such as Atremis) and TweetDeck. Two, material may be passed on to CTIRU from other police and partner agencies, and organisations such as Site Intel and Intel Centre. And three, the final approach is a public referral system in which members of the public can submit a short form identifying the website or app where the material was found and what is the nature of the concern (UK Government 2018). This form is similar in required content to those used by the large platforms. Every public referral

made, whether to a large Internet platform or CTIRU, is reviewed by a human moderator. During and after a terrorist attack, online material is particularly prevalent and there are higher quantities of material to assess.

HOW IS MATERIAL ASSESSED AND REMOVED?

Material is assessed in relation to the policies (or legislation) of a given organisation. For CTIRU, material is assessed in accordance with the Terrorism Act 2006. Material that breaches the Terrorism Act 2006 is detailed carefully, according to those breaches. For instance, a video that breaches the Act would be assessed and evaluated in terms of its content and its location. Content that glorifies or encourages terrorism or terrorists, or that is produced by proscribed terrorist organisations, will be marked according to specific instances of glorification or encouragement, for instance, as evidence of that breach. Timestamps and descriptions of the breaching content are included in a report. Where material is hosted in a UK jurisdiction, it is a legal requirement for the platform hosting the material to remove it. However, much of this material is typically hosted outside of UK jurisdiction, particularly in the USA, where First Amendment and differential norms on Freedom of Speech make removal of certain types of material more difficult. In these cases, CTIRU will contact platforms (which are not limited to those mentioned here) and request that the platform removes the material. Whilst such requests include information of breaches to the Terrorism Act 2006, it is through assessment of the material in accordance with the platforms' own policies or terms and conditions that the bulk of the request is made. Interestingly, the larger platforms' policies and explanations of what is and is not acceptable can be more proscriptive – and therefore more powerful – than the Terrorism Act 2006 in assessing online terrorist material. Nevertheless, the final decision as to whether an item is removed remains with the platform in question. Organisations such as CTIRU and EU IRU have 'Trusted Flagger' status with some of these major platforms (Europol 2016b). This essentially means that any material flagged by these units will be queued as high priority

and will be addressed first by the organisations' content moderators. Good relations between law enforcement and providers and platforms are critical for the removal of terrorist material in this case (Europol 2016a).

According to Gillespie, the way that content moderation is conducted by large platforms is a critical part of shaping user participation and occurs via "[moderation] (through removal, filtering, and suspension); they recommend (through news feeds, trending lists, and personalised suggestions); and they curate (through featured content and front-page offerings)" (2018, p. 202). However, there is not a great deal of information available as to how the material is assessed or the way that platforms make their decisions. The way in which a conduct moderator plays his or her role remains unclear. Roberts suggests that content moderation is:

> *Almost always done in secret for low wages by relatively low-status workers, who must review, day in and day out, digital content that may be pornographic, violent, disturbing, or disgusting. The workers act as digital gatekeepers for a platform, company, brand, or site, deciding what content will make it to the platform and what content will remain there.*
>
> **2016, p. 147**

Indeed, a recent exposé by the Dispatches team on Channel 4 revealed a host of failings in the way in which Facebook moderators were trained, and issues with the policies guiding decisions (Dispatches 2018a). Prior to Mark Zuckerberg's pledge to enhance his security team to 20,000 employees, 7,500 staff were employed as content reviewers for Facebook around the world, with much of this work being outsourced (Bickert 2018; Dispatches 2018a). Whilst these moderators might have relevant language skills, there is generally little subject expertise, and a short, three-and-a-half-week training period revolves around learning about Facebook policies and how to implement them (Bickert 2018). Like CTIRU and EU IRU, these

moderators are available all day, every day to review and remove content (YouTube 2018b). Like CTIRU case officers, who are well trained over a long period of time and are typically recruited from within the police force, these content moderators assess the material they receive according to the policies they have guiding them. For instance, whilst it is acceptable for video material to show a person dying, it is not permissible for that death to be due to dismemberment. ISIS videos showing executions by beheading would therefore easily be removed. However, material that represents extreme political views – which is most of the content that is flagged to Facebook – will not be removed and, at most, will be Marked As Disturbing (Dispatches 2018a). Thus, whilst content from an individual professing support of ISIS or any of ISIS's methods might be removed due to its glorification of terrorism, this decision would rest on the subjective opinion of the moderator reviewing it. Additionally, if the material supported the goals and objectives of ISIS without explicitly mentioning the group or individuals associated with it, then the material would not warrant removal at all, meaning it would remain online with the potential to radicalise sensitive viewers.

It is important to note that Facebook has both frontline and secondary content moderators. The frontline staff are those who are outsourced, but full-time specialist staff support the more difficult decision-making processes (Dispatches 2018a). Little is known about how this section of staff makes decisions as to what material is removed and what material is not, but the Dispatches documentary did reveal some inconsistencies between policy and practice. For instance, some groups and/or individuals that might be classed as extremist, and who repeatedly violate Facebook terms and conditions, are nonetheless shielded from removal, allegedly where these pages have a large number of members (Dispatches 2018a). According to Dispatches, the Britain First Facebook page was marked for eight or nine policy violations, where the number of such violations warranting removal is only five. Whilst Britain First is not a proscribed terrorist group, but a legal political party, it does use right-wing extreme speech (Britain First 2018). Britain First had, until March 2018 when its page was deleted, over two million

followers. In this particular case, content moderators on the frontline were noting these clear policy violations, but these violations were apparently ignored by the full-time Facebook staff.

Indeed, one particularly challenging area for Internet platforms and IRUs is far-right extremist material. Whilst unpleasant and potentially harmful, extremist content is not illegal, yet it is regularly flagged. If an extremist group becomes proscribed as a terrorist group, as National Action was in 2016, only then does any content produced by it or proclamations of support for that group become illegal. Extremist material, and far-right extremism more generally, are grey areas. It is unlikely that extremist material will be classified as terroristic, but may be classified as hate crime, depending on the content and the context. There are policies relating to hate crime for all of the larger platforms, but, for IRUs, the case would be passed along to different departments. If content moderation and AI use existing knowledge to build a schema of what is going to be identified as problematic on the basis of being terroristic (for example, ISIS or al-Qaeda), then being able to identify groups that may be emerging as threatening or extremist, particularly in the context of the far right, is likely to be a challenge given the importance of retaining freedom of speech. This challenge is compounded by the apparent shift to the political right in many Westerns countries. In a context in which right-wing rhetoric (such as anti-Muslim and anti-immigrant sentiment) is increasingly being presented by political leaders,[11] identifying and distinguishing far-right extremism in online material, and being able to justify its removal without impinging on freedom of speech, becomes increasingly difficult. This, of course, has wider implications for those who may be targeted or affected by such content.

11 For instance, Donald Trump recently tweeted that three American congresswomen should "go back and help fix the totally broken and crime infested places from which they came." (www.twitter.com/realDonaldTrump/status/1150381395078000643) [Accessed 23 July, 2019].

CHALLENGES AND CONCERNS

Therein lies the ultimate concern with human assessment and moderating online content: it is built around understanding of the context in which the material arises, and it is subjective, guided by sometimes fairly broad policies, and sometimes, apparently, ignored. Whilst the organisations discussed here are clear about the necessity of human reviewers, the problem of biases – both organisationally and individually – does not seem to be given much consideration. According to YouTube, "Human reviewers remain essential to both removing content and training machine-learning systems because human judgement is critical to making contextualised decisions on content" (YouTube 2017). This judgement can include who has produced the material, the context in which it is being published, and whether it crosses the line from extremist to terrorist. What is also interesting about the way in which terrorist content is reviewed by human moderators is how it typically comes to their attention. Aside from some small law-enforcement units, the content is typically flagged by members of the public, or provider or platform users. Thus, this material is first being assessed by members of the public on some basis. Using the public as the very frontline of content assessment incurs another layer of subjectivity and bias, and possibly makes the process more complex and time-consuming than it may otherwise be.

Despite these limitations and concerns, it is clear that the larger platforms are working towards holding themselves accountable for the content they publish. Although this may be in part due to pressure applied from governments and the public, organisations such as Facebook do seem to be working to adapt to the problems associated with this ever-changing environment. Facebook's turnaround in terms of policies, and new training and hiring, was very quick after the exposé. Furthermore, it – like the other platforms – is committed to freedom of speech, and therefore thinks carefully about what it removes (Dispatches 2018b). The problem is that this thought – and decision-making process is not transparent enough for critical engagement and, therefore, assessment of it. Yet, given the fine,

blurry, sometimes indiscernible line between extreme views and what can also be considered terrorist material, perhaps this lack of transparency should not be surprising.

References

Bickert, M. 2018. 'Content Moderation: Facebook'. *Santa Clara Law*, USA. Retrieved from: www.santaclarauniversity.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=6e2bf22d-52cd-4e3f-9324-a8810187bad7

Britain First. 2018. 'Britain First'. Available at: www.britainfirst.org/ [Accessed 15 July, 2018].

Dispatches. 2018a. 'Inside Facebook: Secrets of a Social Network'. Retrieved from: www.channel4.com/programmes/inside-facebook-secrets-of-a-social-network/on-demand/67241-001 [Accessed 10 July, 2018].

Dispatches. 2018b. Interview with Richard Allen (VP Global Policy Facebook) on 'Inside Facebook: Secrets of a Social Network'. Retrieved from: www.channel4.com/programmes/inside-facebook-secrets-of-a-social-network/on-demand/67241-001 [Accessed 10 July, 2018].

Europol. 2016a. *EU Internet Referral Unit Year One Report*. Retrieved from: www.europol.europa.eu/publications-documents/eu-internet-referral-unit-year-one-report-highlights

Europol. 2016b. 'Europol joins Forces with Counter-Terrorism Experts to Undermine Online Terror'. Available at: www.europol.europa.eu/newsroom/news/europol-joins-forces-counter-terrorism-experts-to-undermine-online-terrorist-propaganda [Accessed 15 July, 2018].

Facebook. 2018a. 'Facebook: Community Standards'. Retrieved from: www.en-gb.facebook.com/communitystandards/ [Accessed 16 July, 2018].

Facebook. 2018b. 'Facebook: How to Report Things'. Available at: www.facebook.com/help/reportlinks [Accessed 16 July, 2018].

Facebook. 2018c. 'Facebook Policies: Dangerous Individuals and Organisations'. Retrieved from: www.en-gb.facebook.com/communitystandards/dangerous_individuals_organizations [Accessed 16 July, 2018].

Gillespie, T. 2018. 'Platforms Are Not Intermediaries'. *Georgetown Law Technology Review* 2 (2), pp. 198–216.

Google. 2018a. 'Google: User Content and Conduct Policy'. Retrieved from: www.google.com/intl/en-US/+/policy/content.html [Accessed 16 July, 2018].

Google. 2018b. 'Report Abuse on Google+'. Available at: www.support.google.com/plus/answer/6320425?hl=en&visit_id=0-636683969253869674-732834091&rd=1 [Accessed 16 July, 2018].

Roberts, S.T. 2016. 'Commercial Content Moderation: Digital Laborers' Dirty Work'. *Media Studies Publications* 12.

Terrorism Act 2006. 2006. London, UK. Available at: www.legislation.gov.uk/ukpga/2006/11/contents [Accessed 02 July, 2016].

Twitter. 2018a. 'Twitter: Report Violations' Retrieved from: www.help.twitter.com/en/rules-and-policies/twitter-report-violation [Accessed 16 July, 2018].

Twitter. 2018b. 'Twitter Rules: Violent Extremist Groups'. Retrieved from: www.help.twitter.com/en/rules-and-policies/violent-groups [Accessed 16 July, 2018].

UK Government. 2018. *Report Online Material Promoting Terrorism or Extremism*. London. Retrieved from: www.gov.uk/report-terrorism [Accessed 16 July, 2018].

YouTube. 2017. 'Expanding Our Work Against Abuse of Our Platform'. *Official Blog: Broadcast Yourself*. Retrieved from: www.youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html [Accessed 17 July, 2018].

YouTube. 2018a. 'YouTube: Account Terminations'. Retrieved from: www.support.google.com/youtube/answer/2802168?hl=en [Accessed 16 July, 2018].

YouTube. 2018b. 'YouTube: Policies and Safety'. Retrieved from: www.youtube.com/intl/en-GB/yt/about/policies/#community-guidelines [Accessed 16 July, 2018].

YouTube. 2018c. 'YouTube: Report Inappropriate Content'. Retrieved from: www.support.google.com/youtube/answer/2802027?hl=en [Accessed 16 July, 2018].

YouTube. 2018d. 'YouTube Community Guidelines Strike Basics'. Retrieved from: www.support.google.com/youtube/answer/2802032?hl=en [Accessed 16 July, 2018].

YouTube. 2018e. 'YouTube Policies: Violent or Graphic Content'. Retrieved from: www.support.google.com/youtube/answer/2802008?hl=en-GB [Accessed 16 July, 2018].

REMOVING AND BLOCKING EXTREMIST CONTENT

Valentine Crosset

As we saw in the previous essays, it is clear that extremist and terrorist groups exploit digital communications. However, the response that should be taken and the actors involved are less apparent. Islamic State's effective use of social media – specifically, Twitter and YouTube – has brought this issue into focus, particularly since 2014. The management of extremist content involves a series of actors, including governments and private actors, who are not used to co-operating with one another. This poses a series of challenges regarding the co-ordination, co-operation and appreciation of the illegal nature of the content (Crosset and Dupont 2018). Whilst the removal of an account or content is often a favoured solution, the challenge goes beyond simply 'tracking down' extremists. It touches on an important definitional problem for public and private actors: if the definition of 'online extremism' or 'hate speech' is too broad, it risks censoring valuable and protected political speech, but a definition that is too narrow risks failing to disrupt extreme networks. Moreover, non-violent extremism raises some challenging questions that require further investigation. First, is it worthwhile to consider hate speech and extremism along a continuum in making decisions about the moderation or takedown of content? In many cases, hateful opinions are neither violent nor are they legally circumscribed as 'hate speech'; nevertheless, they can inspire violent action (take Darren Osborne's attack on Finsbury Park Mosque as an example, or Anjem Choudary's extreme interpretation of Islam that inspired the Woolwich attackers). Should content management and regulation consider taking down non-violent extremist content? And what is the relationship between non-violent and violent extreme speech? This section of the report will focus on responses to online extremism formulated across different governments and the private sector, and particularly how it is paired with increasingly demanding laws.

INTERNET INDUSTRIES

In the United States, platform providers are relatively flexible in setting their own content standards. Section 230 of the 1996 Communication Decency Act gives them broad immunity about what users post on their platform. As Gillespie explains, it gives a sort of double immunity to platform providers. The first "ensures that intermediaries that merely provide access to the Internet or other network services cannot be held liable for the speech of their users" (Gillespie 2018, p. 30). The second protects the intermediaries if they choose to moderate content of their users. In a nutshell, "choosing to delete some content does not suddenly turn the intermediary into a 'publisher', nor does it require the service provider to meet any standard of effective policing" (2018, p. 30). However, due to the rapid spreading of Islamic State-related content, the US and other governments began putting pressure on Silicon Valley actors to strengthen their content regulation, and, more specifically, to enhance their automatic detection tools.

The reactions of private actors to the dissemination of extremist and violent content have evolved over time and have differed from one platform to another. For instance, in 2015, a Facebook spokesperson said, "There is no place for terrorists on Facebook" (Greenberg 2015), whilst, to the contrary, Twitter has long maintained one of the most tolerant policies of freedom of expression among the major social networks (Ducol 2015). Nevertheless, due to various governmental pressures, digital platforms have all strengthened their fight against extremist content, primarily through content moderation. In addition to having tightened their usage policies, over the past few years, web giants have reinforced their moderation teams (thousands of employees), developed automated tools to detect extremist content (particularly via artificial intelligence techniques), and co-operated with other digital platforms to share information and best practices in order to better identify extremist content. The most notable of such collaborations is the Global Internet Forum to Counter Terrorism (GIFCT). Announced in 2016, this partnership between

Google, Microsoft, Facebook and Twitter uses a database of unique fingerprints of visual content to identify and take down that which has been determined to be extremist (Microsoft 2017).

Initially, content reporting was done mainly through users and governments; however, automated techniques have become the primary intermediary and have allowed platforms to censor a priori (MSI-NET 2016). For example, Facebook (2018) indicates that 99% of deleted al-Qaeda- and Islamic State- related content is detected automatically, before any human reporting. For Twitter (2017a), 95% of accounts suspended for promoting terrorism are the result of proprietary Twitter tools (compared to 74% in the 2016 transparency report). Of these, 75% were suspended prior to posting their first tweet. In total, from August 2015 to April 2018, Twitter removed 1.2 million accounts promoting terrorism. As for YouTube, more than half of the videos that were removed for violent extremism were taken down before a human being had even reported them.

The practice of moderation poses considerable challenges to firms. For private actors, it is difficult to assess terrorist content, which some believe should be a judge's responsibility (Hecker 2015). Another significant challenge is the fact that images disseminated by terrorist groups do not necessarily violate terms of service. The resulting editorial choices are then "based on deontology or morality, not on law" (Hecker 2015, p. 29). Moreover, because of the nature of the volume and rapid evolution of content, it is difficult for the moderation teams to maintain control over the content. Several moderators also expressed their concerns about the inconsistency and peculiar nature of some of the internal rulebooks guiding their work, which launched debate about social media giants' ethics (Hopkins 2017).

LAW ENFORCEMENT AND GOVERNMENTS

Government actions to counter extremist online discourse are varied and multiple. First, they can, as with any user, make requests for removal. Removal requests from government can go through one of two processes. First, by making legal requests to remove or withhold

content. Second, through standard customer-support intake channels to review content against Twitter's terms of service (Twitter 2017b). In the Government TOS reports from July to December 2017, government requests represented less than 0.2% of all suspensions in the reported time period and reflected a 50% reduction in accounts reported as compared to the previous reporting period.

Following an increased terrorism threat level and various attacks in Europe, several European countries have strengthened laws aimed at countering the digital footprint of extremist movements (Ducol 2015), whilst the US has favoured partnerships with companies owning digital platforms (Crosset and Dupont 2018). This can be explained by the fact that the companies are under American law, which limits the power of governments to intervene, as freedom of expression is protected by the First Amendment of the US Constitution. Therefore, censorship is naturally controversial in democratic societies; we can see it is not consistent across states.

In terms of law, there are two regulation techniques in this case: the ones that regulate user behaviour directly; and those that regulate by restricting social media companies and the Internet. For example, in November 2014, France created several legal frameworks that targeted the Internet, including harsher sentencing in the case of promotion of terrorism online; authorisation to administratively block terrorist websites; and the possibility for the administrative authorities to ask Internet service providers (ISPs) to block access to sites that glorify, promote or incite terrorism. These laws have been enormously controversial both legally (breach of the principle of separation of powers) and practically (possibility of circumvention; Audureau and Seelow 2015). In Germany, since 1 January, 2018, through the law called Network Enforcement Act (or 'NetzDG'), social media companies such as Twitter, Facebook, YouTube, Instagram and Snapchat have been forced to remove illegal content from their sites within 24 hours of being notified. If they fail to comply, they face fines of up to €50 million. In the same vein, the European Union has warned tech giants to remove terror content

within one hour of receiving the removal order and has called upon these companies to advance a common set of tools to detect, block and remove terrorist content (Bodoni 2018).

For the first time, these different legislative devices and recommendations hold social media accountable for any hateful content published on their platforms. Citron argues that this will "exacerbate censorship creep" due to "definitional ambiguity, global enforcement of companies' speech rules, and opacity of private speech practices" (Citron 2018, p. 1,051). Whilst Citron identifies important tensions – not least that the use of censorship laws to address terrorism might lead to a problematic effect on other, valuable forms of speech – clarity in identifying 'hate speech' and 'terrorist material' can alleviate this issue (2018).

Finally, the UK government (Home Office 2018) announced in February 2018 that it has created a new technology tool to automatically detect terrorist content on all digital platforms, discussed by Gallacher in this report. According to the governmental release, the model has been trained only on Islamic State videos, which excludes other terrorist content and other forms of extremism. Recent tests have shown that their tool was able to detect 94% of Islamic State material with 99.995% accuracy. Developed by the Home Office and ASI Data Science, the tool uses advanced machine-learning techniques that analyze the sounds and images of a video to determine whether or not it is propaganda. The tool can be used on any platform and integrates into the upload process so that most video propaganda is stopped before it reaches the web. Some questions and criticisms can nevertheless be raised (Temperton 2018). First, some pointed out that the effectiveness of such a system against changes in Islamic State propaganda is questionable. Second, the tool targets only one type of content, namely the official videos produced by Islamic State's central and provincial media teams, which excludes the multitude of other contents. Finally, this tool also poses a problem of implementation as no guarantee exists as to whether the platforms will use it.

CONSEQUENCES

The government co-optation on the giants of the web has the consequence that the latter are increasingly forced to take on the role of gatekeepers. They are asked to take proactive measures – such as the use of automated detection tools – to better protect their platforms and their users from terrorist abuse. Even if some regulations target all digital platforms, so far only the giants of the web have the technological, human and economic resources needed to implement such filtering systems. How will less popular platforms and encrypted messaging adapt to these legislations when they do not have the same means and resources as the tech giants? These are essential questions. Many of the extremist content has migrated to other, less regulated platforms or encrypted messaging (Berger and Perez 2016; Prucha 2016), allowing content to continue to circulate and be archived across multiple platforms.

Some concerns can also be raised about the effectiveness of takedowns. Whilst takedown seems a logical approach to disrupting violent extremist behaviour, there are three further issues that need to be taken into account. First, disruption on Twitter has led to the migration of pro-Islamic State activity to more marginal and private systems such as Telegram, a messaging application (Prucha 2016). This poses challenges for researchers, as well as intelligence agencies and police, as these communications are encrypted and cannot be easily accessed or disrupted. Second, facing suspension on Twitter and other social media platforms can be a badge of pride for extremists and can play a role in community-building among these networks (Pearson 2018). And third, the question of disruption requires circumscription of what actually counts as extremism, a definition that has always been contested. In the context of right-wing extreme digital speech, which often cloaks its extremist beliefs under the veil of rationality and, in public, actively avoids the use of hate speech, it can be much harder to identify violent extremism and its 'non-violent' variants. Further, right-wing extreme digital speech is often dealt

with in the register of hate speech, which causes further hesitation when it comes to the decision on whether to take down a user's account or their content.

References

Audureau, W. and Seelow, S. 2015. 'Les Ratés de la Première Vague de Blocages Administratifs de Sites Jihadistes'. Le Monde, 18 March. Retrieved from: www.lemonde.fr/pixels/article/2015/03/18/les-rates-de-la-premiere-vague-de-blocages-administratifs-de-sites-djihadistes_4596149_4408996.html

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', Program on Extremism occasional paper. George Washington University.

Bodoni, S. 2018. 'EU warns Tech Giants to Remove Terror Content in 1 Hour – or Else'. Bloomberg, 1 March. Retrieved from: www.bloomberg.com/news/articles/2018-03-01/remove-terror-content-in-1-hour-or-else-eu-warns-tech-giants

Citron, D. K. 2018. 'Extremist Speech, Compelled Conformity, and Censorship Creep', SSRN Scholarly Paper No. ID 2941880. Rochester, NY. Social Science Research Network. Retrieved from: www.papers.ssrn.com/abstract=2941880

Crosset, V., and Dupont, B. 2018. 'Internet et Propagande Jihadiste: la Régulation Polycentrique du Cyberespace'. *Critique internationale*, (1), pp. 107–125.

Ducol, B. 2015. 'Comment le Jihadisme est-il Devenu Numérique?'. *Sécurité et Stratégie*, 20 (1), pp. 34–43.

Facebook. 2018. 'Hard Questions: How Effective Is Technology in Keeping Terrorists off Facebook?'. *Newsroom.* Retrieved from: www.newsroom.fb.com/news/2018/04/keeping-terrorists-off-facebook/

Gillespie, T. 2018. 'Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media'. New Haven. Yale University Press.

Greenberg, J. 2015. 'Why Facebook and Twitter Can't Just Wipe Out ISIS Online'. *Wired*, 21 November. Retrieved from: www.wired.com/2015/11/facebook-and-twitter-face-tough-choices-as-isis-exploits-social-media/

Hecker, M. 2015. 'Web Social et Djihadisme: Du Diagnostic aux Remèdes'. *Focus Stratégique*, 57, pp. 1–47. Retrieved from: www.ifri.org/fr/publications/etudes-de-lifri/focus-strategique/web-social-djihadisme-diagnostic-aux-remedes

Home Office. 2018. 'New Technology Revealed to Help Fight Terrorist Content Online'. Retrieved from: www.gov.uk/government/news/new-technology-revealed-to-help-fight-terrorist-content-online

Hopkins, N. 2017. 'Revealed: Facebook's Internal Rulebook on Sex, Terrorism and Violence' *The Guardian*, 21 May. Retrieved from: www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence

Microsoft. 2017. 'Facebook, Microsoft, Twitter and YouTube announce formation of the Global Internet Forum to Counter Terrorism', Microsoft blog. Retrieved from: www.blogs.microsoft.com/on-the-issues/2017/06/26/facebook-microsoft-twitter-youtube-announce-formation-global-internet-forum-counter-terrorism/

MSI-NET. 2016. 'Study on the Human Rights Dimensions of Automated Data Processing Techniques (in Particular Algorithms) and Possible Regulatory Implications'. Council of Europe. Retrieved from: www.rm.coe.int/study-on-algorithmes-final-version/1680770cbc

Pearson, E. 2018. 'Online as the New Frontline: Affect, Gender, and ISIS-Take-Down on Social Media'. *Studies in Conflict and Terrorism*, 41(11), pp.850–874.

Prucha, N. 2016. 'IS and the Jihadist Information Highway – Projecting Influence and Religious Identity via Telegram'. Perspectives on Terrorism, 10 (6). Retrieved from: www.terrorismanalysts.com/pt/index.php/pot/article/view/556

Temperton, J. 2018. 'ISIS Could Easily Dodge the UK's AI-powered Propaganda Blockade'. *Wired*, 13 February. Retrieved from: www.wired.co.uk/article/isis-propaganda-home-office-algorithm-asi

Twitter. 2017a. 'New Data, New Insights: Twitter's Latest #Transparency Report', 13 February. Retrieved from: www.blog.twitter.com/official/en_us/topics/company/2017/New-Data-Insights-Twitters-Latest-Transparency-Report.html

Twitter. 2017b. 'Government TOS reports – July to December 2017'. Retrieved from: www.transparency.twitter.com/en/gov-tos-reports.html

EVALUATING THE PROMISE OF FORMAL COUNTER-NARRATIVES

Bharath Ganesh

Formal counter-narratives are central to the activities of Countering Violent Extremism (CVE) programmes (Davies et al. 2016, p. 59), despite the fact that they lack a clearly articulated theory and a solid evidence base (Glazzard 2017). The Institute for Strategic Dialogue (ISD) argues that there is a continuum of activities that can fall under the broader 'spectrum' of counter-narrative work, including 'counter-narratives', 'alternative narratives' and 'government strategic communications' (Glazzard 2017, p. 7; Briggs and Feve 2013). This essay chooses to focus on the term 'counter-narrative' to identify specific programmes that are centred on addressing and challenging extremist narratives. Other formulations, such as 'counter-messaging', can often overlap with counter-narratives. This overlap can lead to inconsistencies that might stunt the efficacy of these programmes by focusing more on the form of messaging rather than the content and meaning of a narrative (Glazzard 2017, p. 7). This essay provides a brief review of key literature on formal counter-narratives. First, it sets out the context and definition of counter-narratives, and then reviews some of the research on counter-narrative campaigns. Running through a few examples from across Europe and elsewhere, and focusing specifically on work done by ISD and Moonshot CVE, the essay provides an up-to-date review of existing approaches in this field. The essay argues that whilst counter-narratives have an important role to play in countering extremist ideology, they are ultimately limited tools to counter EDS.

Often, counter-narrative campaigns involve actors across government, the private sector and civil society. A growing body of research has demonstrated that terrorist narratives are an important tool for recruitment, and has consequently proposed counter-narratives that are "designed to contradict the themes that fuel and sustain terrorist

narratives, and by extension, discourage the support for terrorism they foster" (Braddock and Horgan 2016, p. 382). Counter-narratives should be differentiated from 'alternative narratives' which present a "positive story about social values, tolerance, openness, freedom and democracy" and 'strategic communications' from governments that are used to correct false information and develop relationships with key constituents (Briggs and Feve 2013; ISD and RAN CoE 2015).

There is very little research on the effects of counter-narrative campaigns, despite the attention the strategy receives from international and national actors, technology companies, and civil society. Whereas content regulation and censorship are seen as 'negative' approaches to countering extremism online, strategic communication and counter-narratives are seen as 'positive' approaches that have more potential to deter extremism and bypass the difficult challenges about content regulation and censorship (Stevens and Neumann 2009). Whilst there is extensive research that contributes insights from a variety of disciplines on how counter-narrative campaigns should be structured (Bertram 2016; Beutel et al. 2016; Braddock and Horgan 2016; Cherney 2016; Braddock and Morrison 2018), the literature lacks systematic evaluation of counter-narrative campaigns (Davies et al. 2016; Gielen 2017). Whilst this literature is growing and it is not possible to provide a comprehensive review here (Gielen 2017; Schmitt et al. 2018), these studies do indicate that counter-narrative programmes are not ideally positioned to combat violent extremism. This is partly due to issues in the evaluation of counter-narrative approaches, which lack data about audiences beyond metrics (often provided by the online platforms from which they are disseminated), as well as the targeting of specific communities; some identify groups that are 'at risk', whilst others are targeted more generally.

For example, McDowell-Smith et al. (2017) judge the efficacy of counter-narrative content based on focus groups with college students, but this is not the primary audience that theorists of counter-narratives identify as groups to target. Whilst it is extraordinarily difficult to measure the efficacy of counter-narratives of potential terrorists, reliance on simple metrics such as impressions and clicks on Facebook or tests on audiences that are not central to those

groups that violent extremists seek to recruit does not provide much detail on how counter-narratives might actually be accomplishing the goal that has been set out for them: to debunk and deconstruct the narratives of violent extremists. Further, as evidenced in a review of three Danish counter-narrative programmes, the approach taken incorporates too wide an audience (often broadcasting to the many to reach the few) and, consequently, can be counter-productive, given that those who seek out violent extremist narratives are likely already sympathetic to such views (Hemmingsen and Ingrid Castro 2017, p. 29). This is compounded by the fact that right-wing extremists tend to use online forums to network virtually (as described by Ganesh in this report), which are more closed spaces that are more difficult to infiltrate than a mainstream platform such as Facebook. This is also the case with jihadis, whose use of Telegram ensures a space for their communications that is more enclosed and difficult to access (as described by Gluck and Bindner in this report). Further, it is clear that violent activity is more likely connected with personal interaction, rather than solely virtual ties (Bigo et al. 2014; see also Ahmed, in this report), which casts further doubt that counter-narratives in the social media space will have desired effects.

In a review of six CVE programmes that involved counter-narratives, Davies et al. (2016) find that they generally failed to incorporate the theoretical approaches proposed by CVE scholarship. In a recent review of research on CVE initiatives, Gielen (2017) identifies only six reports on the efficacy of counter-narrative campaigns out of seventy-three total reports on CVE campaigns. This dearth of data analysis is symptomatic of a lack of primary empirical analysis across terrorism studies that Gill et al. identify (2017, p. 103). The EU-wide Radicalisation Awareness Network (RAN) project reports on seventeen projects undertaken in Europe (and a few from abroad). These projects feature a variety of activities across the counter-narrative spectrum, but very few provide information and facts that actively debunk terrorist narratives; rather, it appears that they have mostly been used to encourage young people and social media users to counter hate speech rather than violence. Thus it might be concluded that counter-narratives are

more useful in reducing extremism and hate, rather than the use of violence. The majority of the projects' report evaluations are based on statistics from content on Facebook pages – primarily impressions, views and clicks (Brown and Marway 2018, pp. 11–56). In another instance of counter-narratives disseminated in Iraq, Speckhard et al. use these metrics alongside content to understand efficacy (Speckhard et al. 2018). Such metrics do not indicate whether or not these initiatives have been successful, nor do they provide detail on behavioural change.

A recent review of interventions by ISD is particularly useful in digging into this question on behavioural change (Davey et al. 2018, p. 13). ISD used Facebook to initiate conversations with targeted individuals who engaged with pages associated with far-right and jihadist extremism, and who used hateful, extreme and dehumanising language. They found that 21% of those associated with Islamist or jihadist extremism responded to counter-narrative interventions, and 16% of those associated with the far right responded to outreach from intervention providers. Of these, a majority did sustain engagement with an intervention provider. And of those who did engage with intervention providers, approximately one in ten contained indicators of potential positive impact in terms of behavioural change, defined in their report as responses which include "suggestions that a conversation may have changed an individual's mind, admission that their online behaviour may be harmful to others, or requests to continue a conversation on another medium" (ibid., p.16). Whilst it is encouraging that a sliver of extremist audiences online can be engaged in this way, ISD's most recent work in this area suggests that counter-narratives are limited in facilitating positive behavioural changes.

The 'Redirect Method' pioneered by Moonshot CVE is a well-known example of another counter-narrative programming approach whose principles have been used across the world. The idea behind the Redirect Method is to produce a 'living document' that is actively updated with expert research on indicators of violent extremist ideology. Then, content is produced and advertising software is used to target those exposed to extremist content online

with counter-content that debunks or falsifies extremist content (Brown and Marway 2018). Whilst the Redirect Method's approach is likely to identify the correct audiences and push appropriate counter-narratives at those most at risk, a study by Schmitt et al. (2018) on counter-messaging on YouTube demonstrates that such an approach of linking counter-messaging content algorithmically to extremist content can fail to achieve its goals, concluding: "algorithmic linkage to extremist content could contribute to polarization processes [e.g. Bright 2018] and even make a positive effect of [counter-messaging] more unlikely" (2018, p. 801). Thus, the selection of audiences is extremely important in counter-narrative work, and the potential effects of increasing polarisation must be considered prior to their deployment.

Whilst there is a serious lack of research into counter-narratives, evaluation of these programmes thus far indicates that their efficacy is not commensurate with the confidence placed in their potential. First, there is an over-reliance on quantitative metrics of viewership that do not indicate whether relevant audiences were reached or any disengagement from terrorist views actually occurred (see Helmus and Klein, 2018). Second, audiences are not being targeted or specified as judiciously as CVE scholars have recommended; consequently, better training for counter-narrative initiatives is necessary. Third, evidence suggests that formal counter-narratives may even have counter-productive effects.

Though it is admirable that many counter-narrative initiatives across the world have incorporated civil society organisations and that technology companies are supporting this work, the confidence in counter-narratives (as they are being used today) is excessive. Counter-narratives are expedient insofar as they allow practitioners, states and social media platforms to avoid difficult debates about censorship and free speech, but the relatively surface-level evaluations of the efficacy of these campaigns do not tell us much about whether they are successful in turning extremists away from the narratives that support violence.

The evidence, however, does show that counter-narratives might be more promising in the area of countering hate speech, rather than violent extremism. For example, EdVenture's Peer to Peer programme (a global initiative) has shown that its campaigns have encouraged university students to actively speak out against hate speech online. Similarly, Hope Not Hate, a British not-for-profit, provides facts, handbooks and resources that investigate hate and extremism and provide tools for civil society to detect and counter extremism (Brown and Marway 2018, p. 15). These two initiatives are just a few among others that tackle hate, which is particularly relevant to countering right-wing extremism and debunking its narratives. However, there is limited information on these initiatives' effects in defusing extremist narratives amongst this audience. Also, given that Davey et al. (2018) find that those participating in far-right extremist discourse online are less likely than jihadists to engage with counter-narrative initiatives, even in this area, counter-narrative efforts do not seem to have a high efficacy.

At best, counter-narratives are likely to be useful, but limited, tools in challenging extremist narratives salient in an audience, rather than effective ones in disengaging violent extremists. Counter-narratives should be understood as a supplementary activity in the broader response to online extremism by government, civil society and the private sector. In terms of combating Islamic extremism, counter-narratives are less likely to disrupt the radicalisation process and should not be a primary pursuit. All the same, counter-narratives should not be abandoned, as challenging extremist narratives is crucial to a broader counter-extremism strategy. These initiatives should be focused on targeting specific audiences, particularly with input from and in partnership with Muslim communities (when they are the target audiences). This could provide members of faith communities with resources to debunk terrorist narratives, but counter-narratives should not be expected to disengage and deradicalise those already enculturated in violent extremist communities online. Counter-narrative programmes may have more potential to counter hate speech. Whilst counter-narratives seem like an effective way to counter violent extremism and the narratives

that support it, the evidence that they deradicalise individuals enculturated into these worldviews is thin. Counter-narratives should continue to be pursued, as these programmes are in their early iterations, but expectation that these are cost-effective counter measures to violent extremism is misplaced. They should be pursued alongside more direct intervention with extremists and existing law enforcement procedures.

References

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), pp. 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).

Bigo, D., Bonelli, L., Guittet, E. P. and Ragazzi, F. 2014. *Preventing and Countering Youth Radicalisation in the EU*. Brussels: European Parliament, Directorate General for Internal Policies. Retrieved from: www.europarl.europa.eu/RegData/etudes/etudes/join/2014/509977/IPOL-LIBE_ET(2014)509977_EN.pdf

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

Braddock, K. and Morrison, J. F. 2018. 'Cultivating Trust and Perceptions of Source Credibility in Online Counternarratives Intended to Reduce Support for Terrorism'. *Studies in Conflict and Terrorism*. pp.1–25.

Briggs, R. and Feve, S. 2013. 'Review of Programs to Counter Narratives of Violent Extremism: What Works and What Are the Implications for Government?'. 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), pp. 17–33.

Brown, K. and Marway, H. 2018. 'Preventing Radicalisation to Terrorism and Violent Extremism: Delivering Counter – or Alternative Narratives'. Radicalisation Awareness Network.

Cherney, A. 2016. 'Designing and Implementing Programmes to Tackle Radicalization and Violent Extremism: Lessons from Criminology'. *Dynamics of Asymmetric Conflict*, 9 (1–3), pp. 82–94.

Davey, J., Birdwell, J. and Skellett, R. 2018. 'Counter-Conversations: A Model for Direct Engagement with Individuals Showing Signs of Radicalisation Online'. Institute for Strategic Dialogue. Retrieved from: www.isdglobal.org/isd-publications/counter-conversations-a-model-for-direct-engagement-with-individuals-showing-signs-of-radicalisation-online/

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), pp. 51–86.

Gielen, A. J. 2017. 'Countering Violent Extremism: A Realist Review for Assessing What Works, for Whom, in What Circumstances, and How?'. *Terrorism and Political Violence*, pp. 1–19.

Glazzard, A. 2017. 'Losing the Plot: Narrative, Counter-Narrative and Violent Extremism'. International Centre for Counter-Terrorism. The Hague. Retrieved from: www.icct.nl/wp-content/uploads/2017/05/ICCT-Glazzard-Losing-the-Plot-May-2017.pdf

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

Helmus, T.C. and Klein, K., 2018. Assessing Outcomes of Online Campaigns Countering Violent Extremism. Retrieved from: www.averagemohamed.com/wp-content/uploads/2019/01/RAND_RR2813.pdf

Hemmingsen, A. S. and Ingrid Castro, K. 2017. *The Trouble with Counter-Narratives*, No. DIIS Report 2017: 1. Copenhagen, DIIS – Danish Institute for International Studies.

Institute for Strategic Dialogue and RAN Centre of Excellence. 2015. RAN Issue Paper: *Counter Narratives and Alternative Narratives.* Radicalisation Awareness Network.

McDowell-Smith, A., Speckhard, A. and Yayla, A. S. 2017. 'Beating ISIS in the Digital Space: Focus Testing ISIS Defector Counter-Narrative Videos with American College Students'. *Journal for Deradicalization*, (10), pp. 50–76.

Schmitt, J. B., Rieger, D., Rutkowski, O. and Ernst, J. 2018. 'Counter-messages as Prevention or Promotion of Extremism?! The Potential Role of YouTube Recommendation Algorithms'. *Journal of Communication*, 68 (4), pp. 780–808.

Speckhard, A., Shajkovci, A., Wooster, C. and Izadi, N. 2018. 'Mounting a Facebook Brand Awareness and Safety Ad Campaign to Break the ISIS Brand in Iraq'. *Perspectives on Terrorism*, 12 (3), pp. 50–66.

Stevens, T. and Neumann, P. 2009. Countering Online Radicalisation: *A Strategy for Action*. The International Centre for the Study of Radicalisation and Political Violence.

INFORMAL COUNTER-NARRATIVES

Kate Coyer

WHY INFORMAL NARRATIVES?

'Counter-narratives' has become a sweeping term for a broad array of activities, ranging from grassroots initiatives and civil-society campaigns to corporate responses and governmental programs. The Institute for Strategic Dialogue (ISD) defines counter-narratives responding to extremism as messages that offer "a positive alternative to extremist propaganda, or alternatively aim to deconstruct or de-legitimise extremist narratives" (Tuck and Silverman 2016, p. 41). Some strategies seek to debunk popular extremist views; others aim to put forth an alternative (counter) narrative. Counter-speech itself is as varied as the kinds of extremism it responds to.

Counter-messaging is a core component of current counter-terrorism strategies, yet these initiatives are not without controversy and critique (Lee 2019a). Counter-narratives have long been questioned over their impact and efficacy, which is largely because most initiatives are backed by governments, which are sometimes not seen as legitimate or trustworthy voices (Lee 2019a). As a result, most research examines formal counter-narrative campaigns, typically undertaken by governments or established non-governmental organisations, especially those civic projects funded by state agencies that require evaluation and impact assessment under the terms of funding (Lee 2019a). However, there are also less informal interventions led by individuals and/or small groups all over the world creating content online that is critical of extremism. These informal actors and actions are potentially more credible and impactful in delivering counter-messages, though less is known about them. One study, for example, examined the significance of Indonesian women informally challenging extremism and their efficacy by flying under

the radar (True and Eddyono 2017). Much of what is formally called counter-messaging is likely to occur in more interpersonal and private settings as part of everyday life – for example, informal conversations with friends and family. More research is emerging that examines the contributions of friends and family in preventing violent extremism (ibid.).

This essay explores informal counter-narratives. In doing so, it looks more at the sources of counter-narratives and not just the narratives themselves. The first part of the essay offers examples of informal counter-narratives and considers the role of humour and parody. This is not to make light of the issues and harms, but rather reflects some of the organic ways in which we communicate and cope. The second part of the essay explores the limits of counter-narratives within the context of informality, including why governments aren't the best messengers. The chapter concludes by considering some avenues for collaboration and impact measurement.

CIVIC INITIATIVES IN COUNTER-MESSAGING

An ISD report (Tuck and Silverman 2016) examines a range of civic initiatives, highlighting different strategies and tactics that organisations have taken. Examples include: positive stories from alternative perspectives (EXIT USA, a project of Life After Hate, run by former white supremacists); highlights of how extremist activities have negatively impacted the people they claim to represent (Global Survivors Network (GSN)); exposing factual inaccuracies, hypocrisies or lies propagated and 'setting the record straight' (Sabahi and Magharebia platforms); demonstrating hypocrisy by highlighting examples of how actions by extremists can be inconsistent with their expressed beliefs (*Average Mohamed* series of short animated videos). These initiatives have a range of financial and organisational support, sometimes including support from the US Department of Defense,

as is the case with the Sabahi and Magharebia platforms. There are handbooks for helping civic groups engage in counter-speech online and develop counter-narrative campaigns.[12]

In many cases, the most effective counter-messages are likely to come from individuals' immediate social environments, including friends and family. The most visible form of informal counter-messaging typically comes in the form of media content, often produced for reasons other than explicit counter-messaging. Many examples of informal counter-messaging use satire or mockery to undermine the credibility of extremist messages. One powerful critique of the extreme right to emerge in recent years is the widespread viral success of the 'Muslamic Ray Guns EDL Anthem'. A (possibly) drunk supporter of the British far-right group English Defence League (EDL) was interviewed on television. When asked why he was there, the EDL supporter offered an incomprehensibly slurred response that likely was 'Islamic rape gangs' but was heard as 'Muslamic ray guns'. This was then turned into a hook for an autotuned song, crafted by Alex Vargas and uploaded to YouTube,[13] where it attracted over 1.9 million views, and gave rise to the 'Muslamic Ray Guns' meme reproduced on T-shirts and elsewhere.

The far-right English Defence League was also targeted by the English Disco Lovers (EDL), a group formed by four individuals who launched a 'Google bomb' campaign using search engine optimisation (SEO) to get their name and content to appear higher in search results and on social media than the 'other' EDL (Tuck and Silverman 2016). With the slogan 'Don't hate, gyrate', the group played on their disco name as they twisted the vocabulary of a hate group and in their words "turn the tables in favour of equality and respect" (Lynskey 2013; par3). Another example is the 2010 comedy film Four Lions that portrays an inept band of would-be suicide bombers in the North of England patterned after the 2005 7/7 attackers.

12 See, for example, ISD's website, www.counternarratives.org

13 www.youtube.com/watch?v=AIPD8qHhtVUandt=22s

CULTURE JAMMING, PARODY, AND HUMOUR

'Culture jamming' represents one way in which existing media and cultural products are themselves transformed into commentary, often through parody and satire (Meikle 2007). Mina, for example, explores Chinese meme culture as a vehicle for socio-political critique within the context of the country's intense censorial regime (2014). As a tactic, it comes out of social movement activism and is not often the framework used to describe counter-narrative efforts. However, it is a useful way to understand some of the informal initiatives that seek to disrupt or subvert mainstream messages and repurpose them into something else. Put more broadly, 'remix culture' is the societal space in which derivative works that combine existing images, audio or text into new products are encouraged (Lessig 2008) and enabled by the democratising potential of digital technologies that make remix so spreadable (Jenkins et al. 2013; Lessig 2008).

Remix culture and jamming doesn't have to be complicated – simple actions can have strong visual impact or satirical bite. A video of American white supremacist Richard Spencer getting punched in the face by an anti-fascist protestor whilst being interviewed went viral on YouTube. Its visceral image created heated public debate over whether or not such an action was justifiable or defensible. When the clip was remixed to a song by the band Devo, it garnered tens of thousands more views, using pop culture to juxtapose the real violence at the heart of white supremacist ideology against the street violence of throwing a punch to demonstrate their nonequivalence. Open the website 'Can I punch Nazis?' and the message on the webpage reads: "Yes, it is always OK to punch a Nazi", with a link to an in-depth Talking Points Memo explainer. The site's clever URL – www.canipunchnazis.com – plays off the debate to drive home its message through a simple website with high search results.

There are also many small content producers active on social media and in other spaces creating parodies critical of extremist groups. These include social media accounts designed to parody extremist content such as the Twitter account 'Britain First Logic' (@britfirstposts), or the Subreddit Behold the Master Race sites in

which posters share examples that demonstrate the profound flaws in white supremacist logic (r/beholdthemasterrace). In other examples, content producers are more directly linked to their advocacy work – for example, the Facebook pages Muslims Against ISIS and Preston Against Hate. ISIS Karaoke was a short-lived Twitter account that received mainstream press attention for its clever juxtaposition of images of ISIS fighters presumably found online with incongruous lyrics from popular songs. The premise is very simple, and the account only ever posted around thirty tweets, but the images themselves serve to undermine ISIS's own imagery through disarming their mythology of power and fear. Still images of ISIS combatants holding weapons juxtaposed with well-known pop lyrics that are trivial or humorous have the impact of turning bravado into farce.[14]

Using mockery to satirise a terrorist organisation is tricky. Making light of brutality is challenging and often fails by either not being funny or missing the mark by relying on dangerous anti-Islamic rhetoric (Dowling 2015). Parody might inadvertently draw attention to hateful people and acts, or serve to normalise hate. Satire and irony sometimes get twisted into thinly veiled justifications for hate speech: when white supremacist Richard Spencer was called out for saying 'Heil Hitler' to a crowd in Washington DC who responded with the Nazi salute, he claimed he was being ironic, despite evidence to the contrary.[15] The extreme right has been successful in weaponising irony to reach their target audiences (Greene 2019). However, when used effectively and appropriately, its disarming quality can help messages reach new and wider audiences and can root them in more familiar cultural contexts. An ISD report on the far right concluded that counter-narratives "must penetrate alternative platforms and burst extreme-right bubbles with campaigns that build on a thorough understanding of Internet culture and counter-cultures" (Davey and Ebner 2017, p. 6).

14 www.twitter.com/_isiskaraoke

15 www.theatlantic.com/politics/archive/2016/11/richard-spencer-speech-npi/508379/

Meme culture makes the viral spread and online dissemination not only easier but more integrated into everyday life and social media spaces. Humour and parody are of course not the only forms of spreadable content, nor are they intended to undermine the seriousness of what is at stake and the violent and dangerous reality of extremism. However, the objectives of informal counter-speech are often to raise awareness among the wider public of hypocrisies or fallacies of extremisms, contradict misinformation from extremist groups, disarm dangerous speech, express anger or concern, or simply engage in current affairs through the means they have available to them – rendering humour and parody powerful tools in an individual's arsenal.[16]

LIMITS OF COUNTER-NARRATIVES

A leading criticism of counter-messaging, apart from insufficient evidence of its impact and effectiveness, has been the credibility of the organisations producing and disseminating messages. Little is known about many of these informal groups or their exact motivation. Less still is known about the audiences for their content or their impact. Despite the lack of evidence base, it is important that informal producers of counter-messaging content be taken into consideration alongside formal producers. Examples such as the Redirect Method suggest that informally produced content is likely to have a greater role in future organisation of counter-messaging campaigns, including grassroots content amplified through their methodology.

Government-backed interventions, in particular, have been criticised. In some instances, attempts at covert influence have been 'exposed' or uncovered by the media (Cobain et al. 2016). Even where governments have sought overtly to enlist the help of local civil society groups, they have been seen as tainting their efforts (Herd and Aldis 2006). The issue of credibility has been addressed by some larger CVE organisations that have sought

16 See Benjamin 2018 for qualitative study of informal counter speech initiatives in the UK.

to make greater use of 'natural world' counter-messaging content, promoting external material in their campaigns rather than creating their own (Silverman et al. 2016, p. 9). This insight builds on the recognition that not all counter-messaging campaigns can or should stem from either the government or civil society organisations. This is why informal responses, even small-scale ones, have import.

Little is known or understood about how audiences access and engage with content. This issue is not resolved through informal narratives, but we must be ever mindful that counter-messaging campaigns are not unleashed into a void. Audiences are already potentially exposed to anti-extremist messages from sources very different from those undertaking formal counter-messaging campaigns (Lee 2019a). It is often unclear whom counter-narrative programs are targeting, what the theory of change is and how to measure the impact of these programs. There is little evidence of the effectiveness of counter-narratives in either preventing violent extremism or deterring recruitment (Lindekilde 2012; Fink, Romaniuk and Barakat 2013; Ferguson 2016; Mastroe and Szmania 2016).

CAN IMPACT OR EFFICACY BE MEASURED?

The concept of counter-narratives as a way to push back against extremist propaganda, recruitment and conspiracy theories has been well established; what isn't known, however, is whether or not it is effective. In practice, it has proven difficult to curate narratives and counter-speech campaigns in a systematic way that has targeted it towards the intended at-risk audiences, as well as to measure impact on behaviour and attitudes. At the same time, for better or worse, counter-narratives are a substantial focus of CVE efforts; serious money is being poured into these efforts, which can provide an alternative to content removal by trying to create spaces for plurality of debate and information access.

The assumption underlying counter-messaging programs is that offering an alternative set of facts or interpretations, debunking myths and exposing lies will alter people's attitudes and, eventually, impact behaviours. These are not unlike strategies to combat

dis-information and mis-information, and have long-term impacts that are not easy to measure or quantify. Counter-narratives need to be much more robust and have social mediation behind them, a conclusion from a VOX-Pol workshop organised with Harvard's Berkman Klein Center for Internet and Society and Central European University.[17] A Demos study examined the extent to which different types of counter-speech are produced and shared on Facebook. They found counter-narrative has potential, but it is not as effective as it could be, and that there are some kinds of counter-speech that can actually be counter-productive (Bartlett and Krasodomski-Jones 2016). A more rigorous and evidence-led approach to understanding it is needed. The study found there is no all-encompassing approach, and that counter-speech content operates differently depending on the context in which it is produced, as well as whether or not the audience's interaction follows a terrorist event (2016). The study also found counter-speech using humour received consistently high levels of engagement across different countries. It also concluded that in the UK, mosques and Muslim educational organisations were failing to reach out to young people via social media and recommended relying on more popular content.

Content- and account-removal by social media companies can undermine attempts at publicising legitimate counter-narratives. The citizen journalism group Raqqa is Being Slaughtered Silently (RBSS) exemplifies this, having had its content removed from Facebook, YouTube and Telegram on multiple occasions despite the fact it is seeking to counter Islamic State messages and narratives. Facebook accounts reporting on or documenting what the UN has termed a "textbook case of ethnic cleansing" (Safi 2017, par 1) of the Muslim-minority Rohingya population in Myanmar were also shuttered or removed (Woodruff 2017).

17 The workshop took place October, 2017 at Harvard Law School: www.voxpol.eu/events/8675/

LEGITIMACY, OR WHY GOVERNMENTS AREN'T THE BEST MESSENGERS

The most effective counter-narratives have proven to be those perceived as authentic and produced by members of a given community. Authenticity of message and messenger are crucial. Government motives are rightly criticised and often so tied to foreign policy directives as to render the message being received more akin to propaganda. Governments have sought to work with civil society and the private sector in their CVE programs, though there persists the perception (often, but not always, correctly) that governments are the ultimate drivers of the message. That fear often results in real distrust by the target communities of CVE campaigns, both of the message and intentionality, especially in terms of how their information might be weaponised against them in arrests or prosecutions: "Chief among the concerns has been the capacity and authority of counter-messaging providers to deliver convincing and authoritative counter-messages. Put simply, why would audiences listen to a word the counter-messaging 'industry' has to say?" (Lee 2019b, p. 1).

POTENTIALS AND PARTNERSHIPS

It is in this climate that unaffiliated individuals and small groups of people have sought to engage through modest interventions, often using humour and remix to craft messages and memes within popular cultures. There is potential for future collaboration between formal and informal civic projects, and some success stories as well as crucial caveats (ibid.). These informal actions are certainly not intended to eradicate extremism in and of themselves, but nevertheless can help inoculate publics against the violence and hate of extremist speech and have a role to play in shifting debate and confronting extremist narratives.

References

Bartlett, J. and Krasodomski-Jones, A. 2016. 'Counter-speech on Facebook'. *Demos*. Retrieved from: www.demos.co.uk/project/counter-speech-on-facebook-phase-2/

Cobain, I., Ross, A., Evans, R. and Mahmood, M. 2016. 'Inside Ricu, the Shadowy Propaganda Unit Inspired by the Cold War'. *The Guardian*, 2 May. Retrieved from: www.theguardian.com/politics/2016/may/02/inside-ricu-the-shadowy-propaganda-unit-inspired-by-the-cold-war

Davey, J. and Ebner, J. 2017. 'The Fringe Insurgency: Connectivity, Convergence, and Mainstreaming of the Extreme Right', Institute for Strategic Dialogue.

Dowling, T. 2015. 'Isis Karaoke: Satire's Answer to Hate Preachers with Microphones'. The Guardian, 31 August, Retrieved from: www.theguardian.com/global/2015/aug/31/isis-karaoke-satires-answer-to-hate-preachers-with-microphones

Ferguson, K. 2016. 'Countering Violent Extremism Through Media and Communication Strategies. Partnership for Conflict, Crime and Security Research'. University of East Anglia.

Fink, N.C., Romaniuk, P. and Barakat, R. 2013. 'Evaluating Countering Violent Extremism Programming: Practice and Progress. Government of Canada and Global Counterrorism Forum'. Retrieved from: www.globalcenter.org/wp-content/uploads/2013/07/Fink_Romaniuk_Barakat_EVALUATING-CVE-PROGRAMMING_20132.pdf

Greene, V. 2019. *Studies in American Humor*, 5 (1), pp. 31–69. Penn State University Press.

Herd, G. and Aldis, A. 2006. 'Synthesizing Worldwide Experiences in Countering Ideological Support for Terrorism (CIST)'. In Aldis, A. and Herd, G. (eds) *The Ideological War on Terror: Worldwide Strategies for Counter Terrorism*. Routledge, pp. 245–253.

Jenkins, H., Ford, S. and Green, J. 2013. *Spreadable media: Creating Value and Meaning in A Networked Culture*, New York. New York University Press.

Lee, B. 2019a. 'Informal Countermessaging: The Potential and Perils of Informal Online Countermessaging'. Studies in Conflict and Terrorism, 42 (1–2), pp. 161–177.

Lee, B. 2019b. 'Countering Violent Extremism Online: The Experiences of Informal Counter-messaging Actors'. *Policy and Internet.*

Lessig, L. 2008. *Remix: Making Art and Commerce Thrive in a Hybrid Economy*. Penguin.

Lindekilde, L. 2012. 'Value for Money? Problems of Impact Assessment of Counter-Radicalisation Policies on End Target Groups: The Case of Denmark'. *European Journal on Criminal Policy and Research*, 18 (4), pp. 385–402.

Lynskey, D. 2013. 'How to Disco Dance the EDL off Google and Facebook'. *The Guardian*, 1 February. Retrieved from: www.theguardian.com/technology/shortcuts/2013/feb/01/disco-dance-edl-google-facebook

Mastroe, C. and Szmania, S. 2016. 'Surveying CVE Metrics in Prevention, Disengagement and De-Radicalization Programs'. START (National Consortium for the Study of Terrorism and Responses to Terrorism). Retrieved from: www.start.umd.edu/publication/surveying-cve-metrics-prevention-disengagement-and-de-radicalization-programs

Meikle, G. 2007. 'Stop Signs: an Introduction to Culture Jamming'. In Coyer, K., Dowmunt, T. and Fountain, A. (Eds.), *The Alternative Media Handbook*, pp. 166–179. London. Routledge, Taylor and Francis Group.

Mina, A.X. 2014. 'Batman, Pandaman and the Blind Man: A Case Study in Social Change Memes and Internet Censorship in China'. *Journal of Visual Culture*, 13 (3), pp. 359–375.

Safi, M. 2017. 'Myanmar Treatment of Rohingya Looks Like "Textbook Ethnic Cleansing", says UN'. *The Guardian*, 11 September. Retrieved from: www.theguardian.com/world/2017/sep/11/un-myanmars-treatment-of-rohingya-textbook-example-of-ethnic-cleansing

Silverman, T., Stewart, C.J., Amanullah, Z. and Birdwell, J. 2016. 'The Impact of Counter-Narratives: Insights from a Year-Long Cross-Platform Pilot Study of Counter-Narrative Curation, Targeting, Evaluation and Impact'. Institute for Strategic Dialogue. Retrieved from: www.isdglobal.org/wp-content/uploads/2016/08/Impact-of-Counter-Narratives_ONLINE.pdf

True, J. and Eddyono, S. 2017. 'Preventing Violent Extremism: Gender Perspectives and Women's Roles'. Monash University. Retrieved from: www.monash.figshare.com/articles/Preventing_Conflict_and_Countering_Violent_Extremism_through_Women_s_Empowerment_and_Civil_Society_Mobilisation_in_Indonesia/7100873

Tuck, H. and Silverman, T. 2016. 'The Counter Narrative Handbook. Institute for Strategic Dialogue'. Retrieved from: www.isdglobal.org/wp-content/uploads/2016/06/Counter-narrative-Handbook_1.pdf

Woodruff, B. 2017. 'Exclusive: Facebook Silences Rohingya Reports of Ethnic Cleansing'. The Daily Beast, 18 September. Retrieved from: www.thedailybeast.com/exclusive-rohingya-activists-say-facebook-silences-them

DIGITAL LITERACY VS THE ANTI-HUMAN MACHINE: A PROXY DEBATE FOR OUR TIMES?

Huw Davies

Sir Tim Berners-Lee, whose HTTP protocol enabled the web, has been recently quoted as saying, "we have demonstrated that the web has failed." Instead of "serving humanity", he said, we now have a "large-scale emergent phenomenon, which is anti-human" (Brooker 2018). Given that he has previously argued that the web was a humanistic artefact of the Enlightenment (Berners-Lee et al. 2006) – a 'social machine' (Berners-Lee and Fischetti 1999, p. 172) – we can assume that by 'anti-human', Berners-Lee means the web has become a place where tolerance, rational thought and scientific epistemologies that promote human progress have been overwhelmed by their binary opposites.

This report is about extreme speech, but it is difficult to isolate its exponents from the rest of the anti-human machine. Extreme speech is often codified so that group insiders know what is being said but outsiders cannot identify it as illegal. This includes visual codes in memes and the co-opting of well-known brands to produce a whole iconographic subculture of extremist thought (Miller-Idriss 2018). Extreme speech is also a product of conspiracy theories: people think their views are justified because they are fighting a malevolent hidden power. Extreme speech has to be contextualised within the whole ecology of digital media within which the boundaries of what is acceptable – questioning the numbers murdered in the Holocaust, for example – prepare the territory in advance for extreme speech to flourish. Only some definitions of digital literacy address extreme speech specifically (such as Vaikutyte-Paskauske, Vaiciukynaite and Pocius 2018); because they can't be easily disaggregated from extreme speech, the many techniques and strategies of the anti-human machine, such as disinformation campaigns, are included here within the same equation.

The anti-human machine is a powerful combination of 'the social' (human actors) and 'the technical' (the affordances of digital technologies). The anti-human machine is placing demands on democracies to address ideologies and methods that are undermining their ability to function. This essay explains what the anti-human machine is, how it functions, and why responses to its effects are eluding the web's big platforms (such as Facebook, Twitter and Google) and confounding national governments. It then shows that, although none of its advocates claim it is a unilateral panacea, digital literacy is being offered as an alternative or additional remedy to the anti-human machine (see, for example, European Commission 2019). However, existing digital literacy solutions, particularly in the UK, are inadequate responses to the challenges our democracies now face. Calls for digital literacy have a long way to go before they catch up with political reality.

There are many ways that technologies have been exploited in service of the anti-human machine. Photoshopped images; forged and redacted videos including so-called 'deep fakes' (using simulation technology to create new synthetic content from existing material, for example, to confect a politician's speech); the deliberate misreporting of events on partisan 'news' platforms; magazines and media outlets that cite pseudo-history, pseudo-science and junk research (research published in low-quality journals that have no rigorous peer-review process), and misrepresent genuine research; the use of 'bots' (both automated accounts and humans behaving as bots) to overwhelm social media feeds via comments, shares or replies; dis-information campaigns; personalised, targeted ads with undisclosed funders; recommender systems that disseminate prejudice and propaganda; and many other techniques that are used to leverage network effects to nourish ideologies that produce hate and extremist speech are all now in play. People exploiting technology can be groups of individuals who show each other online that they share grievances or affiliations. They can participate in more organised or co-ordinated groups that have a history offline and have been rejuvenated by the web, such as Stormfront. They may be paid or motivated by state actors and/or may act within swarms that come together to respond to events, memes, postings or tweets only to disperse before the next event

(Ganesh, second essay in this report). Or they can be part of imagined communities that involve temporary, contingent and messy coalitions of all the above. What can be done about this socio-technical threat?

HOW THE ANTI-HUMAN MACHINE IS RESISTING THE PLATFORM INTERVENTION

The major platforms have had to become quasi-governmental gatekeepers of public discourse. However, after a series of scandals (such as Facebook hosting the live feed of the recent gun massacre in Christchurch, New Zealand), their failure to protect their users from extremism has drawn critical attention worldwide. This criticism has intensified because the platforms' gatekeeping systems lack transparency; when they do apply their corporate policies, their decisions to remove content or groups (such as Britain First) can often appear arbitrary. As Crosset (this report) shows, platforms apply the rules inconsistently depending on the country within which they are operating and even their own employees responsible for content mediation find guidelines confusing and contradictory. Aside from their terms and conditions, there is no publicly available methodology to follow their adjudications about what content they find unacceptable and why; precedents are set, then violated; and there is neither an effective way to challenge their decisions, nor a process to hold them to account.

Platforms are also struggling to accommodate the wider political context within which hate speech or extremist speech can be codified and normalised by mainstream populist politicians. So, if they block an extremist, we may legitimately ask why don't the platforms block the mainstream politicians who are amplifying this extremist's message? Given that the major platforms have billions of users continually adding content, even if they had a transparent and publicly accountable methodology for removing violations of their terms and conditions, they still have to rely on algorithms to identify breaches. At such a scale of content generation, even a 1% failure rate lets through too many transgressions for platforms to be able to deal with manually (see Gallacher, this report). Moreover, once it is possible

to understand the logic of these algorithms, they can be gamed. (For example, to an algorithm, depictions of violence could, for example, be made to look like admissible video game footage.)

HOW THE ANTI-HUMAN MACHINE IS DEFEATING GOVERNMENT INTERVENTION

The alternative to the self-regulation for platforms is government intervention. Civil rights lobbyists are anxious about handing over this power to decide what is acceptable on platforms to governments because it can easily be abused. Libertarian groups argue governments removing content is censorship that violates our right to free speech. If such decisions are handed over to the public, in today's political climate, reaching a democratic consensus about what is acceptable to censor on platform will be challenging, if not impossible. If an agreement is accomplished, how do we prevent majoritarianism violating the rights of vulnerable or marginalised groups? If platforms are unwilling or unable to follow guidelines that emerge from this consensus, how do governments enforce them?

The European Commission is currently formulating legislation to give its member states the power to compel platforms to remove extremist content and hate speech and fine them up to 4% of their global revenue if they fail to comply. But it remains to be seen what happens if the platforms refuse to pay these fines or divert some of their billion-dollar reserves and profits to financing legal teams to challenge any rulings in expensive and protracted court proceedings. Therefore, beyond government intervention, fixing the platforms' users through educational programmes to prevent digital technologies being used in anti-human ways appears to be relatively attractive.

THE DIGITAL LITERACY SOLUTION

Digital literacy has a long and complicated genealogy that includes information, computer and media literacy (see Nichols and Stornaiuolo 2019 for a full discussion). In England, digital literacy is currently delivered to school-aged children through the

national curriculum in computer science (UK Parliament Science and Technology Select Committee 2016). The curriculum focuses on 'up-skilling' target populations, equipping them for a '21st century jobs market', making liberal use of verbs such as 'thrive' and 'participate' at a "level suitable for the future workplace and as active participants in a digital world" (Department of Education 2013).

However, recent discussions within government show that many stakeholders believe this form of digital literacy is no longer adequate. For example, the 5Rights Framework cited by the UK's House of Lords Communications Select Committee (2017) on digital skills states that, via schools, digital literacy should help children and young people "critically understand the structures and syntax of the digital world", "to be confident in managing new social norms" and understand "issues of data harvesting and impact of persuasive design strategies" (Kidron, Evans and Afia 2018). And, following a report from the National Literacy Trust, the recent UK All-Party Parliamentary Group on Literacy stressed that children and young people need to be taught the 'critical literacy skills' to identify 'fake news' (National Literacy Trust 2018). After its investigation into disinformation and fake news, the UK government's Department for Media, Culture and Sport (DCMS) went further, concluding, "digital literacy should be the fourth pillar of education, alongside reading, writing and maths" and delivered "as part of the Physical, Social, Health and Economic curriculum" (2019).

There is another discussion we should be having about establishing an evidence base for digital literacy, including whom to target and why (especially within a profoundly unequal educational system) and what to do about people who are unwilling or unable to access formal education programmes. However, the focus here is on the mismatch between the digital literacy solution as proposed in the policy circles described above and the challenge of the anti-human machine. Whilst research into this area is nascent, the anti-human machine's users are in many ways already digitally literate. They have an acute understanding of the 'syntax of the digital world' and 'persuasive design strategies'. Indeed, these strategies, together with effective reputation management, and, to further quote the 5Rights Framework,

"the confidence to manage new social norms" (Kidron, Evans and Afia 2018), have enabled extremists to reach and mobilise a wider, global audience online. Such malign actors have developed techniques of attention-hacking to increase the visibility of their ideas through the strategic use of social media, memes and automated bots – as well as by targeting journalists, bloggers and influencers to help disseminate content (Marwick and Lewis 2017). As already mentioned, many extremists, who know their views are normatively transgressive, offensive or illegal, have adapted to codify their language and normalise their discourse by successfully crossing the boundaries between marginal and mainstream media, including effectively manipulating the affordances and weaknesses of the platforms. Stormfront's 'style guide', for example, is "particularly interested in ways to lend the site's hyperbolic racial invective a facade of credibility and good faith" (Feinburg 2017). Anti-Semitism, white supremacism, Islamophobia and misogyny are often perpetuated through irony and an intimate knowledge of Internet culture (ibid.). Jihadists such as Islamic State (IS) have also successfully exploited social media platforms – often by mimicking the production techniques and action tropes of Hollywood blockbusters, and they have even produced jihadist computer games (Atwan 2019).

The antidote to this may be more critical thinking, but prominent members of the 'intellectual dark web' community have consciously co-opted the norms and language of academia into their strategies to create a parallel criticality. Critical thinking now means profoundly different things to different people. Adherents of the 'intellectual dark web' already believe they have reached the apogee of their form of critical thinking about the digital and broadcast media, politics and science. It is, perhaps, telling that sociologist Bruno Latour has recently been reflecting ruefully on his critique of science (Vrieze 2017) because, in its disingenuous application, it has empowered conspiracy theorists and climate change deniers who argue the scientific research on climate change is politically motivated and compromised by its 'biased' sources of funding. The digital ecosystem within which all these ideologically affiliated users and

groups operate on websites such as Reddit, 4chan and 8chan provide learning opportunities
for young people who are drawn to these figures and their ideas.

In response, Emejulu and McGregor (2016) argue digital literacy must confront the politics of the anti-human machine head on. They, therefore, locate digital education within the "wider discursive and material struggles for equality and social justice" (ibid., p. 3) that actively push back against reactionary ideologies. However, there is an obvious danger that such forms of digital literacy, which include online activism, may produce even more polarised and mutually energising antithetical crowds fighting over what is permissible online. We need to ask, how does such a form of digital literacy become part of the solution and not simply another casualty of the so-called online culture wars (boyd 2018)? Given they monetise engagement through advertising and that antagonism boosts engagement, in the culture wars, the platforms are the only winners. Politically engaged users are also sharing sensitive personal data about their views on public platforms for commercial and governmental surveillance agencies to capture.

There is nothing in any existing calls for digital literacy in the policy circles above that addresses the bigger picture here, which is the co-constitutive relationship between psychological mechanisms such as confirmation bias, motivated reasoning and cultural cognition; the heuristics and techniques we use to confirm our in-group status, including aggression towards the Other; the evolving social norms that define how we engage with each other online; the history and ideologies of racism and misogyny, including their theological origins and the tactics of populism and extremism; the ideologies and political economy of platform capitalism; and the deliberate exploitation of ignorance of all of the above.
This means the problem is much bigger than skills. Many of us are not interested in fact-checking or the effort of critical thinking if we are rewarded for being seen to be endorsing disinformation on social media. This produces correlations as people align their views across unrelated domains to conform to the prejudices of their ideological in-group. For example, high levels of racial resentment are strongly

correlated with reduced agreement with the scientific consensus on climate change (Benegal 2018). As a result, we live in a society that is often grossly misinformed, as very few people go away to check the facts on emotive issues. For example, we hugely over-estimate the proportion of Muslims in Britain – we think 21% British people are Muslims when the actual figure is 5%. And we think 24% of the population are immigrants – which is nearly twice the real figure of 13% (Ipsos MORI 2017). Therefore no digital literacy programme is ever likely to work unless it produces reflexive critical thinkers, motivated to challenge their own thinking and positionality: people who know and care when they are being sold a biased or racist view of history or pseudo-science, or when they are being manipulated. As boyd (2018) identifies, digital literacy needs to be about epistemology: how do we know what the facts are and where do we go to find them? It also needs to be about the methods that support independent thinking, understanding claims, and validating knowledge without having to rely on appeals to authority.

It is therefore easy to fall into the trap of digital literacy inflation, where we call for ever-more sophisticated forms of digital literacy that eventually become a whole multi-disciplinary curriculum that we call education. However, the values and practices that should be the foundation of such an education are now being framed within anti-human machine (and beyond) as those of an ideologically perverse, smug, self-serving, distant, liberal or left-wing academic elite unable or unwilling to address the concerns of 'the people' including what Kaufman (2018) calls "white ethnic loss". Given how the web, with the techniques described above, is being used to undermine social cohesion and our collective capacity to address climate breakdown, digital literacy has become a site for a proxy debate for one the most important challenges of our time: how do we rescue knowledge from the anti-human machine?

References

Atwan, A.B. 2019. *Islamic State: The Digital Caliphate*. University of California Press.

Benegal, S.D. 2018. 'The Spillover of Race and Racial Attitudes into Public Opinion about Climate Change'. *Environmental Politics* 27 (4), pp. 733–756.

Berners-Lee, T., Hall, W., Hendler, J. A., O'Hara, K., Shadbolt, N. and Weitzner, D. J. 2006. *Foundations and Trends in Web Science* 1 (1) pp. 1–130.

Berners-Lee, T., and Fischetti, M. 1999. *Weaving the Web*. Orion Business. London.

boyd, d. 2018. 'You Think You Want Media Literacy… Do You?'. Retrieved from: www.points.datasociety.net/you-think-you-want-media-literacy-do-you-7cad6af18ec2

Brooker, K. 2018. 'I was Devastated: Tim Berners-Lee, the Man who Created the Web, has Some Regrets'. *Vanity Fair*. Retrieved from: www.vanityfair.com/news/2018/07/the-man-who-created-the-world-wide-web-has-some-regrets

Emejulu, A. and McGregor, C. 2016. 'Towards a Radical Digital Citizenship in Digital Education'. *Critical Studies in Education* 60 (1), pp. 1–17.

Feinburg, A. 2017. 'This is the Daily Stormer's Playbook'. The Huffington Post. Retrieved from: www.huffingtonpost.co.uk/entry/daily-stormer-nazi-style-guide_n_5a2ece19e4b0ce3b344492f2

House of Lords Select Committee. 2017. 'Interactions with the Digital World'. Retrieved from: www.publications.parliament.uk/pa/ld201617/ldselect/ldcomuni/130/13007.htm

House of Commons Technology and Science Select Committee. 2016. 'Digital Skills in Schools'. Retrieved from: www.publications.parliament.uk/pa/cm201617/cmselect/cmsctech/270/27006.htm

IPOS Mori. 2017. 'Perils of Perception'. Retrieved from: www.ipsos.com/ipsos-mori/en-uk/perils-perception-2017

Kaufman, E. 2018. 'White Majorities Feel Threatened in an Age of Mass Migration – and Calling Them Racist Won't Help'. Retrieved from: www.newstatesman.com/politics/uk/2018/10/white-majorities-feel-threatened-age-mass-migration-and-calling-them-racist-won

Kidron, B., Evans, A. and Afia, J. 2018. 'Disrupted Childhood: The Cost of Persuasive Design'. Retrieved from: www.5rightsframework.com/static/5Rights-Disrupted-Childhood.pdf

Marwick, A. and Lewis, B. 2017. Media Manipulation and Disinformation Online. Data and Society. Retrieved from: www.datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf

Miller-Idriss, C. 2018. 'What Makes a Symbol Far Right? Co-opted and Missed Meanings in Far-Right Iconography'. In Fielitz, M., Thurston, N. (Eds.), *Post-Digital Cultures of the Far Right*, pp. 123–136.

Nichols P. and Stornaiuolo, A. 2019. *Media and Communication* (ISSN: 2183–2439) 2019, 7 (2), pp. 14–24.

The Department for Education. 2013. National Curriculum in England: Computer Science Programmes of Study. Retrieved from: www.gov.uk/government/publications/national-curriculum-in-england-science-programmes-of-study

The Department for Media Culture and Sport. 2019. 'Disinformation and "Fake News": Interim Report'. Retrieved from: www.publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/363/36310.htm#_idTextAnchor061

The European Commission. 2019. European Media Literacy Event. 'Internetkunskap (Internet Knowledge) – Media Literacy and Digital Skills for Adult Citizens'. Retrieved from: www.ec.europa.eu/futurium/en/european-media-literacy-events/internetkunskap-internet-knowledge-media-literacy-and-digital-skills

The National Literacy Trust. 2018. 'Fake News and Critical Literacy: Final Report'. Retrieved from: www.literacytrust.org.uk/documents/1722/Fake_news_and_critical_literacy_-_final_report.pdf

Vaikutytė-Paškauskė J., Vaičiukynaitė J. and Pocius, D. 2018. 'Research for CULT Committee – Digital Skills in the 21st century'. European Parliament, Policy Department for Structural and Cohesion Policies. Brussels. Retrieved from: www.europarl.europa.eu/RegData/etudes/STUD/2018/617495/IPOL_STU(2018)617495_EN.pdf

Vrieze, J. 2017. 'Bruno Latour, a Veteran of the "Science Wars", Has a New Mission". Retrieved from: www.sciencemag.org/news/2017/10/bruno-latour-veteran-science-wars-has-new-mission

The VOX-Pol Network of Excellence (NoE) is a European Union Framework Programme 7 (FP7)-funded academic research network focused on researching the prevalence, contours, functions, and impacts of Violent Online Political Extremism and responses to it.

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 312827

Email info@voxpol.eu
Twitter @VOX_Pol
www.voxpol.eu