NETWORK OF EXCELLENCE FOR
RESEARCH IN VIOLENT ONLINE
POLITICAL EXTREMISM

VOX
Pol

Forums
Announcements
aftermath of a suspected car bomb
Subscribe
Share this video
http://youtu.be/38eYEzsD3LA
5 views
Upload
Rebels Taking Hits From The Army
Hard clashes Rebels caught on camera
Civil War Rages On Heavy Clashes With Rebels
Heavy Clashes As The Army Still Advances
Rebels Taking Hits From The Army
Car Bomb December 2012)
Army Forces In Heavy Combat With Rebels
Rebels Defending
firing improvised cannons

VOX-Pol WORKSHOP ON COMBATING ONLINE EXTREMISM: STATE, PRIVATE SECTOR, AND CIVIL SOCIETY RESPONSES DISCUSSION DOCUMENT

10 – 11 May 2018

www.voxpol.eu

VOX-Pol Workshop ‘Combating Online Extremism: State, Private Sector, And Civil Society Responses’

Discussion Document

10 – 11 May 2018
University of Oxford

TABLE OF CONTENTS

INTRODUCTION

This short discussion paper outlines a set of issues in three thematic areas that we wish to explore in this workshop and the report, the primary output of the workshop. It sets out three themes for research: content regulation and censorship, the role and politics of the private sector in CVE, and counternarratives and the engagement of civil society.

The paper is just a draft that tries to set out a few challenges present in the literature and debate on each of these themes. We request that participants in the workshop take a look at these challenges and consider how their research/presentation might contribute to helping to address the challenges identified. Of course, this document and its claims are tentative and through discussion over the workshop, we hope to introduce more nuance and clarify the challenges that we might address.

CONTENT REGULATION AND CENSORSHIP

For discussion in this theme, we propose four main challenges for future research:

- Measurement of the efficacy of takedowns and other disruption techniques;
- Navigating the controversial politics of censorship that are inevitably raised when engaging in takedown;
- Moving beyond a legal definition of "extremism" and "hate speech" and drawing on interdisciplinary work that understands how these activities are locally situated and defy universal definitions;
- The extent to which content moderation and censorship should apply to non-violent forms of extremist expression.

It is clear that digital communications play a clear role in opening pathways to radicalisation (Edwards & Gribbon, 2013; Gill et al., 2017). However, the response that should be taken, and the actors involved, is less clear. ISIS's effective use of social media—specifically Twitter and YouTube—brought this issue into focus particularly in 2014 and after. Of all forms of response, 'takedown' or the removal of an account or content from a social media platform has been the most controversial. However, as Conway et. al. find, the effect of takedowns of pro-IS accounts by Twitter has "severerly affected IS's ability to develop and maintain robust and influential communities on Twitter" (Berger & Perez, 2016; Conway et al., 2017). They also point out that this focus on IS has led to less pressure on other jihadist groups, making the case that a more diverse range of jihadist groups ought to be considered in the analysis of key actors supporting violent extremism online (Conway et al. 2017, 39).

While 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-ISIS 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, having faced suspension on Twitter and other social media platforms can be a badge of pride for extremists and plays a role in community-building among these networks (Pearson, 2017). There are similar trends for the extreme-right, for whom account suspension is only "proof" that the "globalist elite" that moderates content is fundamentally biased against "conservative" opinions. Third, the question of disruption requires circumscription of what actually *counts* as extremism, a definition that is always contested. In the context of the extreme-right, which often cloaks its extremist beliefs in the veneer 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, extreme-right speech online is often dealt with in the register of "hate speech", which causes further hesitation when it comes to the decision to take down a user's account or their content.

While censorship is naturally controversial in democratic societies, this is not consistent across nation-states. For example, Danielle Citron notes that American "free speech values guided policy decisions in Silicon Valley" (Citron, 2018, 1036), but in the context of extremism and the terrorist attacks in Europe and the consequent pressure placed on social media platforms by European governments, these companies agreed to remove hateful speech within 24 hours in an agreement with the European Union (2018, 1038). Citron argues that this will "exacerbate censorship creep" due to "definitional ambiguity, global enforcement of companies' speech rules, and opacity of private speech practices" (2018, 1051). While Citron identifies important tensions—not least that the use of censorship laws to address a problem like terrorism might lead to a chilling effect on other, valuable forms of speech—clarity in identifying "hate speech" and "terrorist material" can alleviate this issue (2018, 1062).

The debate on "extremism" and "hate speech" has primarily been undertaken through the legal discipline, which risks "universalising normative frameworks" and obscuring more "situated understanding of the cultures of communication and online practices that have been obfuscated by the overarching category of hate speech" (Pohjonen & Udupa, 2017, 1186). Some critical attention is warranted in understanding the situated forms and histories that drive extreme or hateful communication practices that explores their use. Pohjonen and Udupa's (2016) intervention that calls for an anthropological approach that pays closer attention to how particular practices of "extreme speech" are grounded in specific cultures challenges the notion that "extremism" and "hate speech" can easily be detected in the context of social media. An example is the hashtag "#whitegenocide" (the idea that there is a conspiracy to replace whites in Western countries, particularly by Muslims) works as a key refrain for opening pathways to radicalisation by extreme right-wing groups in the United States (Berger, 2016). The term itself does not directly identify extremism or hate speech, but it plays a role in the cultural imaginaries that influence right-wing extremists in the North America and Western Europe. Pohjonen and Udupa (2016) make a convincing argument that universal applications of normative categories such as "hate speech" or "extremism" are questionable when applied across different nation-states or cultural contexts.

It is also clear that non-violent extremists make use of the Internet and it plays a role in the motivation for terrorism. This broaches an important definitional problem: if a 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. Non-violent extremism raises some sticky questions that require further interrogation. First, is it worthwhile to think of a continuum of hate speech and extremism in making decisions about the moderation or takedown of content? In many cases, hateful opinions are not *violent* nor are they legally circumscribed as "hate speech," but they can inspire violent action (take Darren Osborne's attack on Finsbury Park Mosque as an example, or Anjem Choudhary's extreme interpretation of Islam that inspired the Woolwich attackers). Social media, in many cases, has accelerated the movement from exposure to purportedly non-violent forms of extremism and the radicalisation of individuals who later went on to engage in political violence. Should content management and regulation consider taking down non-violent extremism, and how does non-violent extreme speech online connect with its more violent forms?

Though there has been relatively more theoretical work in this area (particularly in discussions of "extremism" and "hate speech"), there are also a few less-developed area. Further research could extend existing research on recommender systems (O'Callaghan, Greene, Conway, Carthy, & Cunningham, 2015) and contagion dynamics (Ferrara, 2017) could help to understand the effects of content moderation and takedown as a disruption to violent extremist activity online. Much of the detection of extremism is being automated, though the criticisms raised above (such as the definitions of "extremism" and "hate speech" and their application in different contexts, as well as the extent to which speech is violent or non-violent) suggest that there may be significant challenges to applying such methods to detect extremists.

THE ROLE (AND POLITICS) OF THE PRIVATE SECTOR IN ONLINE CVE

For discussion in this theme, we propose four main challenges for future research:

- The effectiveness of emerging responses from the private sector and public-private partnerships in delivering CVE
- Critical analysis of the logic and execution of CVE-related tools and initiatives from the private sector
- The diversity of voices involved in shaping how the private sector engages with CVE
- The role of the scientific community in scrutiny of private sector CVE-related practices

The previous section focused on content moderation and censorship from a conceptual perspective. In practice, however, this activity is ultimately the responsibility of private sector companies, albeit executed often following pressure from government or civil society. Recently numerous initiatives have been developed to counter online extremism. These initiatives have often been funded by technology companies including Google, Microsoft, Facebook, and Twitter. Most notable in this area is the recent collaboration announced in 2016, Global Internet Forum to Counter Terrorism (GIFCT). Its primary resource is a database of unique fingerprints of visual content that can be used to identify and takedown content that has been identified as extremist (Microsoft, 2017).

There are a number of other attempts to address extremism, hate, and related problems (eg. misinformation and propaganda) being undertaken by the private sector. For example, Moonshot CVE's "Redirect Method" tries to intervene at the level of online recommender systems (funded by Jigsaw, a division of Alphabet). Google's Perspective API to assist moderators of online discussions to judge the toxicity of user-generated contributions.[1] More recently, the UK Home Office partnered with ASI Data Science to develop a tool to detect IS propaganda online with a high level of accuracy (Greenfield, 2018). Other programmes, such as the Online Civil Courage Initiative (OCCI) is a project developed in partnership between Facebook and the Institute for Strategic Dialogue to support and develop skills to counter online extremism from civil society partners.[2] While these initiatives have admirable goals, it is

[1] See https://www.perspectiveapi.com/#/.
[2] See https://www.isdglobal.org/programmes/communications-technology/online-civil-courage-initiative-2/.

difficult to identify the extent to which such programmes are actually effective in reducing the potential that the Internet has for extremist recruitment.

While these initiatives driven by the private sector are encouraging, there is a paucity of academic literature that critically interrogates the logic and execution of these initiatives. A growing body of research has clearly demonstrated the role that bias plays in machine learning algorithms and it is unclear whether academics and civil society are involved in the training and development of these systems. As has been noted extensively by researchers on political violence, the lack of transparency due to security issues prevents a diversity of members of the scientific community being involved in the development of these systems. With automated forms of content regulation supported by government and in use at companies such as YouTube, without transparency, it will be extremely difficult to understand the efficacy and equality of the outcomes of the use of these systems. Further, it is becoming clear that these systems are designed with militant Islamic political violence (and specifically IS) in mind. Whether these systems and the companies developing them have a blind spot when it comes to extreme right-wing propaganda and violence must be questioned.

There are also serious political concerns that need to be broached when it comes to the access and power that technology companies have (as opposed to civil society, for example) in directly working with governments in developing counter-extremist policy for the Internet. While it is encouraging that these companies have used their substantial resources to address CVE in the online space and in outreach to the UN and national governments, it is not clear if this has excluded the voices of civil society and academia in this agenda. For example, in an open letter published in *Internet Policy Review*, academics point out that Facebook's new initiative to engage with selected groups of social scientists might lead to only specific disciplines and research teams having the ability to engage with the company's data, while others are left out. They note:

> Had Facebook and Twitter listened to scholarly concerns about undifferentiated third-party data access, political bots, and 'fake news', for instance, they could already have acted on these issues well before the political upheavals of 2016…if such research is locked out as a result of the coming API change, all that will remain is the shallow, commercially focussed analysis provided by the major market research companies that are strategic partners of or commercially dependent on Facebook and Twitter - this is neither in the interest of the users, nor ultimately good for the platforms themselves (Bruns, 2018)

While the letter refers to the context of political communication rather than extremism, the points raised are germane to emerging responses from these companies to CVE – if their methods end up becoming black-boxed and prevent scrutiny from the scientific community, or they only engage with a select set of partners, the challenges social researchers have identified previously about the transparency of data relating to terrorism and extremism may be reproduced. Consequently, it is imperative that researchers in this field take a critical stance to the politics in which knowledge about online extremism is produced.

COUNTERNARRATIVES AND THE ENGAGEMENT OF CIVIL SOCIETY

For discussion in this theme, we propose two main challenges for future research:

- Measurement of the effectiveness of counternarrative campaigns;
- Engagement of strategic communications actors with a wide range of stakeholders, particularly civil society;
- The role of trust, particularly amongst minority communities, in the effectiveness and delivery of strategic communications programmes related to CVE.
- Evaluating the limits and potential of “positive” versus “negative” responses to violent online extremism.

Counternarratives are a form of strategic communication that can be used to “contradict the themes that fuel and sustain terrorist narratives” (Braddock & Horgan, 2016). There are numerous examples of counternarrative strategies in use and is one of the rapidly growing areas in current online CVE research. Where content regulation and censorship are seen as ‘negative’ approaches to countering extremism online, strategic communication and counternarratives are seen as ‘positive’ approaches that have more potential to deter extremism and bypassing the difficult challenges about content regulation and censorship (Stevens & Neumann, 2009).

While there is an extensive body of literature that explores the best practices for strategic communication, there is relatively less focus on the effectiveness of these programmes in deterring extremists. The challenges of evaluation and measurement are present in a recent study by Speckhard et al. (2018) note that counter narrative interventions can initiate important conversations with radicalised individuals, but it is relatively less clear whether these interventions lead to disengagement with extremism. Despite the paucity of research on the outcomes of such approaches, there has been a proliferation of research on guiding principles for counternarratives (Beutel et al., 2016), the role of legitimate interlocutors in the dissemination of counternarratives (Braddock & Horgan, 2016), and the importance of using social media as a means towards deradicalisation (Bertram, 2016). These initiatives are relatively recent and, given the long term process involved with deradicalization, it may be too demanding to expect rigorous, large-scale studies that can shed light on the efficacy of counter narratives.

What is clear is the importance of the engagement with a wide range of stakeholders, particularly leaders from faith communities (Mirahmadi, 2016). This kind of consultation can work in concert with law enforcement efforts (Cohen, 2016), but it is important to remember that trust in institutions (particularly from communities that are maligned by CVE programmes) is important to the effectiveness of such programmes. Consequently, it is extremely important that researchers focus on the effectiveness and design of CVE programmes as well as the ways in which they incorporate the voices of a variety of stakeholders.

The emerging responses to online CVE suggest that this kind of engagement may face limits, in particular, the concerns raised in the previous section on the selective ways in which technology companies reach out to community groups affect the diversity of voices that are involved, and can limit the legitimacy that these programmes have. The Prevent strategy in the UK has caused severe distrust amongst British Muslims to some of its initiatives due to its perception that it renders Muslims into a "suspect community" (Awan, 2012; Ragazzi, 2016). This presents significant challenges to community outreach, though these hurdles have been surmounted in various ways. The question that remains, however, is the extent to which online CVE programmes have effectively incorporated voices and handled this issue of trust to result in effective programmes. Further, in terms of right-wing extremism, it is necessary to identify those who play an important role in these communities and can assist in the development of counternarratives

Civil society plays an important role in bridging gaps between policies and the communities in which extremists live and conduct their recruitment. It is demonstrably clear that strategic communications have an important role to play but more discussion is required to understand the limits of strategic communication and its outcomes, as well as the limits and outcomes of content regulation and censorship.

BIBLIOGRAPHY

Awan, I. (2012). “I Am a Muslim Not an Extremist”: How the Prevent Strategy Has Constructed a “Suspect” Community. *Politics & Policy*, *40*(6), 1158–1185. https://doi.org/10.1111/j.1747-1346.2012.00397.x

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.

Berger, J. M., & 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.

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

Beutel, A., Weine, S., Saeed, A., Mihajlovic, A., Stone, A., Beahrs, J., & Shanfield, S. (2016). Guiding Principles for Countering and Displacing Extremist Narratives. *Contemporary Voices: St Andrews Journal of International Relations*, 7(3). https://doi.org/10.15664/jtr.1220

Braddock, K., & Horgan, J. (2016). Towards a Guide for Constructing and Disseminating Counternarratives to Reduce Support for Terrorism. *Studies in Conflict & Terrorism*, *39*(5), 381–404. https://doi.org/10.1080/1057610X.2015.1116277

Bruns, A. (2018). Facebook shuts the gate after the horse has bolted, and hurts real research in the process. Retrieved May 1, 2018, from https://policyreview.info/articles/news/facebook-shuts-gate-after-horse-has-bolted-and-hurts-real-research-process/786

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 https://papers.ssrn.com/abstract=2941880

Cohen, J. D. (2016). The Next Generation of Government CVE Strategies at Home: Expanding Opportunities for Intervention. *The ANNALS of the American Academy of Political and Social Science*, *668*(1), 118–128. https://doi.org/10.1177/0002716216669933

Conway, M., Khawaja, M., Lakhani, S., Reffin, J., Robertson, A., & Weir, D. (2017, August 15). Disrupting Daesh: measuring takedown of online terrorist material and it's impacts

[Monograph]. Retrieved May 1, 2018, from http://www.voxpol.eu/download/vox-pol_publication/DCUJ5528-Disrupting-DAESH-1706-WEB-v2.pdf

Edwards, C., & Gribbon, L. (2013). Pathways to Violent Extremism in the Digital Era. *The RUSI Journal*, *158*(5), 40–47. https://doi.org/10.1080/03071847.2013.847714

Ferrara, E. (2017). *Contagion Dynamics of Extremist Propaganda in Social Networks* (SSRN Scholarly Paper No. ID 2982259). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2982259

Gill, P., Corner, E., Conway, M., Thornton, A., Bloom, M., & Horgan, J. (2017). Terrorist Use of the Internet by the Numbers. *Criminology &Public Policy*, *16*(1), 99–117. https://doi.org/10.1111/1745-9133.12249

Greenfield, P. (2018, February 13). Home Office unveils AI program to tackle Isis online propaganda. *The Guardian*. Retrieved from http://www.theguardian.com/uk-news/2018/feb/13/home-office-unveils-ai-program-to-tackle-isis-online-propaganda

Microsoft. (2017, December 4). Facebook, Microsoft, Twitter and YouTube provide update on Global Internet Forum to Counter Terrorism. Retrieved May 1, 2018, from https://blogs.microsoft.com/on-the-issues/2017/12/04/facebook-microsoft-twitter-and-youtube-provide-update-on-global-internet-forum-to-counter-terrorism/

Mirahmadi, H. (2016). Building Resilience against Violent Extremism: A Community-Based Approach. *The ANNALS of the American Academy of Political and Social Science*, *668*(1), 129–144. https://doi.org/10.1177/0002716216671303

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2015). Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems. *Social Science Computer Review*, *33*(4), 459–478. https://doi.org/10.1177/0894439314555329

Pearson, E. (2017). Online as the New Frontline: Affect, Gender, and ISIS-Take-Down on Social Media. *Studies in Conflict &Terrorism*, *0*(0), 1–25. https://doi.org/10.1080/1057610X.2017.1352280

Pohjonen, M., & Udupa, S. (2017). Extreme Speech Online: An Anthropological Critique of Hate Speech Debates. *International Journal of Communication*, *11*(0), 19.

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

Ragazzi, F. (2016). Suspect community or suspect category? The impact of counter-terrorism as 'policed multiculturalism.' *Journal of Ethnic and Migration Studies*, *42*(5), 724–741. https://doi.org/10.1080/1369183X.2015.1121807

Speckhard, A. (2018). Bringing Down the Digital Caliphate: A Breaking the ISIS Brand Counter-Narratives Intervention with Albanian Speaking Facebook Accounts – ICSVE. Retrieved May 1, 2018, from http://www.icsve.org/research-reports/bringing-down-the-digital-caliphate-a-breaking-the-isis-brand-counter-narratives-intervention-with-albanian-speaking-facebook-accounts/

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

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