Program on Extremism
THE GEORGE WASHINGTON UNIVERSITY

SOCIAL MEDIA, TERRORIST CONTENT PROHIBITIONS AND THE RULE OF LAW

This paper, part of the Legal Perspectives on Tech Series, was commissioned in conjunction with the Congressional Counterterrorism Caucus

STUART MACDONALD

SEPTEMBER 2019

About the Program on Extremism

The Program on Extremism at George Washington University provides analysis on issues related to violent and non-violent extremism. The Program spearheads innovative and thoughtful academic inquiry, producing empirical work that strengthens extremism research as a distinct field of study. The Program aims to develop pragmatic policy solutions that resonate with policymakers, civic leaders, and the general public.

About the Author

Stuart Macdonald's research interests lie in criminal law and counterterrorism, particularly cyberterrorism and terrorists' use of the internet. Macdonald is Director of the University's Cyber Threats Research Centre and a Co-Director of the University's EPSRC-funded £7.6m CHERISH Digital Economy Centre. He is the lead organiser of the biennial Terrorism and Social Media (TASM) conference, a member of Europol's Advisory Network on terrorism and propaganda and co-ordinates Swansea University's contribution to the Global Research Network on Terrorism and Technology. In 2016/17 Macdonald was also the holder of a Fulbright Cyber Security Award.

The views expressed in this paper are solely those of the author, and not necessarily those of the Program on Extremism or the George Washington University.

Introduction

The importance of the rule of law to an effective counterterrorism strategy is widely accepted. Adherence to rule of law values protects both the legitimacy and moral authority of counterterrorism policies and legislation. This paper focuses on two specific rule of law values: minimalism and certainty. Minimalism is concerned with issues of scope. Laws should be as narrowly drawn as possible in order to preserve individuals' autonomy and freedom to choose, to the fullest extent possible. Certainty is concerned with issues of clarity. Laws should be worded as clearly as possible so that individuals are aware of their responsibilities and able to make informed choices about their actions. Narrowly, clearly drawn laws also limit the discretion vested in officials, thus providing protection against inconsistent or inappropriate decision-making by those tasked with implementing the law.

The rule of law is traditionally associated with public institutions, not private technology companies. In the contemporary realm of counterterrorism, however, a steadfast public-private distinction is difficult to maintain. Indeed, many have urged the importance of public-private partnership in responding to terrorists' use of the internet.[1] One specific issue that has generated much discussion has been social media companies' regulation of extremist content on their platforms. Facebook's *Community Standards*, the *Twitter Rules* and YouTube's *Community Guidelines* all expressly prohibit content that promotes terrorism. Most of the discussion of these prohibitions has focused on the speed with which they are enforced, particularly following the attacks in Christchurch, New Zealand.[2] This paper seeks instead to evaluate the prohibitions from the different, but equally important, perspective of the rule of law values of minimalism and certainty.

To inform the discussion, the paper draws on the debates that have surrounded the U.K. 'Encouragement of Terrorism' criminal offence. Created by the Terrorism Act 2006, and recently amended by the Counter-Terrorism and Border Security Act 2019, this offence

has proved controversial from its inception for two principal reasons. [3] [4] [5] [6] First, the offence expressly encompasses both direct and indirect encouragement. Critics have argued that the concept of indirect encouragement is too nebulous and gives the offence too wide a scope. Second, the framing of the offence focuses not on the purpose of the speaker, but on whether the potential effect of the statement is to encourage terrorism. This too, it has been argued, gives the offence too wide a scope.

In terms of the social media companies' prohibitions on terrorism-promoting content, this paper accordingly asks two questions. Do the prohibitions encompass indirect, as well as direct, encouragement? And, for the prohibitions to apply, must the encouragement of terrorism have been the purpose and/or the likely effect of the relevant content? The answer to neither question is clear from the wording of the prohibitions themselves. The paper will argue that, in terms of the values of minimalism and certainty, it is important that the answers to both questions are made explicit. It will also suggest how both questions should be answered and provide a proposed reformulation of the social media companies' prohibitions on terrorism-promoting content.

The U.K.'s Encouragement of Terrorism Offence

The U.K.'s Encouragement of Terrorism offence contains three requirements that must be satisfied for a defendant to be liable. First, the defendant must have published a statement or caused another to publish a statement (s 1(2)(a)). A "statement" is defined as a "communication of any description" and includes communications "without words consisting of sounds or images or both" (s 20(6)). "Publishing" is defined in a similarly expansive manner, as "publishing [the statement] in any manner to the public", and expressly includes providing an electronic service "by means of which the public have access to the statement" and "using such a service ... to enable or to facilitate access by the public to [it]" (s 20(4)). The legislation's accompanying explanatory notes explain that Internet Service Providers and website administrators may therefore be regarded as publishing statements on their platforms/websites. There is one restriction, however:

the statement must have been published to the public. The offence does not apply to private communications.

The offence's second requirement focuses on the content of the statement and its likely interpretation. The prosecution must show that the statement was "likely to be understood by a reasonable person as a direct or indirect encouragement or other inducement, to some or all of the members of the public to whom it is published, to the commission, preparation or instigation of acts of terrorism or Convention offences" (s 1(1)). The "public" is defined as the public (or any section thereof) of any part of the UK or of another country, and expressly includes public meetings or gatherings (regardless of whether payment is required to attend) (s 20(3)).

The final requirement is that the defendant *either* intended to (directly or indirectly) encourage members of the public to commit, prepare or instigate acts of terrorism *or* was reckless as to whether the statement would have this effect (s 1(2)(b)). Since proof of recklessness will suffice, there is no requirement to prove a terrorist purpose. There is, however, a defence of non-endorsement. Only available in cases of reckless encouragement, this defence applies where: (a) the statement neither expressed the defendant's views nor had his endorsement; and, (b) in the circumstances it was clear that the statement neither expressed his views nor had his endorsement (s 1(6)).

Facebook, Twitter and YouTube's prohibitions on terrorism-promoting content are considerably pithier. Facebook's *Community Standards* say, "We do not allow content that praises [terrorists organisations or terrorists] or any acts committed by them."[7] The *Twitter Rules* state that "You may not make specific threats of violence or wish for the serious physical harm, death, or disease of an individual or group of people. This includes, but is not limited to, threatening or promoting terrorism." [8] And YouTube's *Community Guidelines* stipulate that "Content intended to praise, promote or aid violent criminal [including terrorist] organisations is not allowed on YouTube."[9] Whilst not suggesting that social media companies' terms of service should contain the same level of detail as criminal legislation, these platforms' prohibitions leave two key

questions unanswered. Do the prohibitions encompass indirect, as well as direct, encouragement? And, for the prohibitions to apply, must the encouragement of terrorism have been the purpose and/or the likely effect of the relevant content?

From a rule of law perspective, it is important that these questions are answered. The right to freedom of speech is critical in the context of counterterrorism. As Barendt has stated, "We can only respond intelligently to undesirable extremist attitudes, and remove or reduce the reasons why they are held, if we allow them, to some extent, to be disseminated."[10] One of the chief criticisms of the U.K.'s Encouragement of Terrorism offence has been that it is overly broad and, as a result, has a chilling effect on free speech.[11] This over-breadth blurs the boundary between, on the one hand, efforts to prosecute those who encourage acts of terrorism and, on the other hand, efforts to respond to the ideological challenge of terrorism. This apparent overlap can render individuals unwilling to participate in Countering Violent Extremism (CVE) programs for fear of criminal prosecution and can aggravate feelings of suspicion and resentment on the basis that such programs are simply a pretext for spying and surveillance.[12] Indeed, one study of Islamic State (IS) Twitter activity found that suspension played an important role in community-building, with the majority of the accounts studied referring to Twitter's use of suspension as a specific tool to persecute Muslims.[13] It is also important that the boundaries of prohibitions on terrorism-promoting content are communicated as clearly as possible.[14] This not only ensures that users are provided with the information needed to understand their rights and responsibilities when using the platform, enabling them to make informed decisions about the content they choose to post, but also restricts and guides the discretion of content moderators. This both limits the risk of inconsistent – or even inappropriate – decision-making and minimizes the potential for "censorship creep."[15]

Definitional Clarity and the Notion of Indirect Encouragement

The U.K.'s Encouragement of Terrorism offence expressly applies to both direct and indirect encouragement. The rationale for including indirect encouragement was so that the offence would also encompass people who "create the climate of hate in which terrorism can more easily flourish."[16] Whilst some commentators have queried the inclusion of more indirect forms of encouragement, such criticisms are misplaced.[17] This is not simply because of the difficulty distinguishing between direct and indirect encouragement (or between different degrees of indirectness), but because such a distinction would be counterproductive. To focus exclusively on statements that encourage via a direct speech act – that is, statements that employ an explicit performative such as "I encourage you to ..." – would so limit the scope of any prohibition on the encouragement of terrorism as to render it practically worthless.[18] Moreover, indirect forms of encouragement are often more persuasive. In the context of asking others to do something for us, indirectness may help to save their face (public image) needs and, in so doing, may also address our own face needs as speakers.[19] Indirectness can also serve to underline common ground between the speaker and the hearer, and/or construct the speaker's identity.[20] These are benefits that are well understood by marketers, advertisers and politicians. Members of Internet Referral Units have also urged the importance of removing "so-called utopian content, that is texts, images, and videos that praise or glorify extremist lifestyles (e.g., showing unrealistically peaceful scenes from bombarded regions)", explaining that "These forms of propaganda may be as dangerous as graphically violent pieces because they might similarly - or possibly more strongly - mobilise people into action."[21]

To illustrate, consider the following statement, from issue one of Al-Qaeda in the Arabian Peninsula (AQAP)'s English-language online magazine *Inspire*:

> To the Muslims in America I have this to say: How can your conscience allow you to live in peaceful co-existence with a nation that is responsible

> for the tyranny and crimes committed against your own brothers and sisters? How can you have your loyalty to a government that is leading the war against Islam and Muslims?

The statement uses a requestive strategy known as "strong hint." The two rhetorical questions hint at two specific requests for action: not to co-exist peacefully with America and not to be loyal to the American Government. The rhetorical essence of the questions comes from the statement's construction of polarised identity groups, emphasising both the bad properties (tyranny) of the out-group (America and the American Government) and its actions (crimes and war leadership) against the in-group ("our own brothers and sisters" and "Islam and Muslims").

Accepting that indirect forms of encouragement should be included raises the question whether the term is too nebulous to communicate its boundaries with sufficient clarity. The term itself is left undefined in the U.K. legislation, with section 1(4) instead instructing fact-finders to consider the contents of a statement and the circumstances and manner of its publication when deciding whether it amounted to encouragement to terrorism. Whilst this seems appropriate – the meaning of any statement depends not only on its specific wording but also the surrounding circumstances, including the broader extra-linguistic context – the need to assess statements on an individualized basis poses a serious challenge to attempts to give advance warning of the statements that, in general terms, will and will not be held to amount to encouragement. The U.K. legislation accordingly seeks to elucidate the meaning of indirect encouragement by offering an illustrative example. Section 1(3) explains that a statement indirectly encourages terrorism if it satisfies two conditions. The first is that it glorifies the commission or preparation of acts of terrorism. Here, glorification means "any form of praise or celebration" (s 20(2)). The glorification can relate to a past or future terrorist act or to acts of terrorism in general (s 1(3)(a)). The second is that the statement is one from which members of the public "could reasonably be expected to infer that what is being glorified is being glorified as conduct that should be emulated by [them] in existing circumstances" (s 1(3)(b)).

Whilst the glorification-plus-emulation example is useful, the U.K. legislation could have gone further in illustrating the meaning of indirect encouragement, in two respects. First, there could also have been an example focused on the denigration of an out-group, as the quote above from AQAP demonstrates. Second, some indication could also have been offered of the ways in which indirect encouragement may be realized, such as via statements of obligation, suggestory formulae and hints (including in the form of rhetorical questions). Together, these would elaborate the meaning of indirect encouragement still further.

Purpose and/or Likely Effect

As explained above, the U.K.'s Encouragement of Terrorism offence focuses on how the statement in question is likely to be understood by "some or all of the members of the public to whom it is published" (s 1(1)). When combined with the express inclusion within the offence of indirect encouragement, this focus on how the statement would be construed (by perhaps only a small minority of its audience) led some to express concern that the offence would apply to statements that express understanding and which, as a result, have the effect of providing encouragement. An example that was discussed during the Parliamentary debates on the legislation was Cherie Blair's comment at a charity event in 2002, referring to Palestinian suicide bombers, that "As long as young people feel they have got no hope but to blow themselves up you are never going to make progress". These concerns about the offence's reach were exacerbated by the breadth of the U.K.'s statutory definition of terrorism, in particular its lack of any exception for just cause.[22] [23] Since the actions of Nelson Mandela in the early 1960s fall within this definition, publishing a statement that celebrates these actions will amount to the (indirect) encouragement of terrorism if some of those to whom it is addressed could reasonably infer from the statement that Mandela's actions are being glorified as conduct that should be emulated by them in existing circumstances.[24]

The underlying difficulty here is that the U.K. offense ignores the fact that persuasion is, by its very nature, a purposive activity. According to speech act theory, a speaker's

intention is the key aspect to consider in determining the intrinsic meaning of a speech act.[25] Yet it is sufficient for the U.K. offense that the statement was likely to have been understood by its audience as (direct or indirect) encouragement to terrorism and that the maker of the statement was reckless as to this possibility. Proof that their purpose in publishing the statement was to encourage terrorism is not necessary. The practical upshot is that, in the U.K., someone may be convicted of encouraging terrorism even if their publication of the statement was, truly speaking, not an act of encouragement at all.

The U.K. Government has responded to criticisms of the breadth of the Encouragement of Terrorism offense by stating that, in practice, prosecutions may only be brought with the consent of the Director of Public Prosecutions (s 19). This reliance on executive discretion has been described by the U.K. Supreme Court as an abdication of legislative responsibility.[26] It delegates to an unelected official the decision whether an activity should be treated as criminal for the purposes of prosecution and "leaves citizens unclear as to whether or not their actions or projected actions are liable to be treated by the prosecution authorities as effectively innocent or criminal."[27] By analogy, broadly worded prohibitions in social media companies' terms of service fail to make clear to users whether content they post will be treated as impermissible and places this decision in the hands of the content moderator responding to the referral/appeal.

To avoid this combination of overly-broad definition and reliance on individual discretion, social media companies' prohibitions on terrorism-promoting content should reflect the nature of persuasion as a purposive activity and be explicitly limited to instances where content is posted with the objective of encouraging terrorism. There are two possible objections to this proposal. The first concerns its practicality. On the biggest platforms, the vast majority of content is removed automatically. Are machines able to discern the purpose for which content is posted? Progress is being made in this respect: Facebook's uses of artificial intelligence already include language understanding (analysing text that has been removed for praising or supporting terrorist

organisations in order to develop text-based signals that can go into machine learning algorithms to detect similar future posts).[28] In any event, in practice most automated decisions to remove content are based on behavioural cues, such as account features (e.g., how long ago it was opened, how often it posts messages) and message behaviour (e.g., including trending or unrelated hashtags). Decisions based on the content of a post still rely heavily on human involvement.[29] And there is evidence to suggest that the purpose for which content is posted is already a key criterion in this context.[30] [31] The second possible objection is that to require a terrorist purpose would unduly limit social media companies' prohibitions on terrorism-promoting content. In response, it should be noted that content that is posted without any intention to promote terrorism may nonetheless fall foul of other prohibitions, such as those aimed at hate speech.[32] More generally, aside from the fact that, as explained above, encouragement is by its very nature a purposive activity, this objection also undervalues the importance to counterterrorism of freedom of speech. Only removing content that is posted with the objective of encouraging terrorism would be consistent with the First Amendment,[33] protect ideological debate and discussion and avoid generating the feelings of suspicion and resentment that have plagued the U.K.'s Encouragement of Terrorism offense.

Conclusion

This paper began by identifying two questions that Facebook, Twitter and YouTube's prohibitions on terrorism-promoting content currently do not answer: do the prohibitions encompass indirect, as well as direct, encouragement? And, for the prohibitions to apply, must the encouragement of terrorism have been the purpose and/or the likely effect of the relevant content? Based on an examination of the U.K.'s Encouragement of Terrorism offence, it has been argued: first, that indirect encouragement should be encompassed but that more is needed to clearly communicate the boundaries of the term indirect encouragement; and, second, that a focus on the likely effect of a statement, instead of the purpose for which it was made, has detrimental consequences for freedom of speech and ideological discussion.

Drawing on this analysis, the paper offers the following reformulation of social media

companies' prohibitions on terrorism-promoting content:

> It is forbidden to post content that purposely encourages terrorism, either directly or indirectly. Indirect encouragement of terrorism includes the use of statements, suggestions or hints (including rhetorical questions) to either (a) denigrate or dehumanise others or (b) glorify past acts of terrorism in such a way as to imply that others should emulate them.

This reformulation would ensure that the ambit of the prohibition on terrorism-promoting content is both carefully circumscribed and clearly communicated to users. Such respect for the rule of law values of certainty and minimalism is of vital importance in counterterrorism.

References

[1] Kavanagh, C., Carr, M., Bosco, F. & Hadley, A. (2017). 'Terrorist Use of the Internet and Cyberspace: Issues and Responses'. In M. Conway, L. Jarvis, O. Lehane, S. Macdonald & L. Nouri (eds.), Terrorists' Use of the Internet: Assessment and Response. IOS Press.
[2] Intelligence and Security Committee of Parliament (2018). The 2017 Attacks: What needs to change? HC 1694. The Stationery Office.
[3] Choudhury, T. (2009). 'The Terrorism Act 2006: Discouraging Terrorism'. In I. Hare & J. Weinstein (eds.), Extreme Speech and Democracy. Oxford University Press.
[4] Hunt, A. (2007). 'Criminal Prohibitions on Direct and Indirect Encouragement of Terrorism'. Criminal Law Review, 441-458.
[5] Marchand, S. A. (2010). 'An Ambiguous Response to a Real Threat: Criminalizing the Glorification of Terrorism in Britain'. George Washington International Law Review, 42, 123-157.
[6] Macdonald, S. & Lorenzo-Dus, N. (forthcoming). 'Purposive and Performative Persuasion: The Linguistic Basis for Criminalising the (Direct and Indirect) Encouragement of Terrorism'.
[7] "Community Standards," accessed June 21, 2019, https://www.facebook.com/communitystandards/.
[8] "The Twitter Rules," accessed June 21, 2019, https://help.twitter.com/en/rules-and-policies/twitter-rules.
[9] "Policies - YouTube," accessed June 21, 2019, https://www.youtube.com/yt/about/policies/#community-guidelines.
[10] Barendt, E. (2009). 'Incitement to, and Glorification of, Terrorism'. In I. Hare & J. Weinstein (eds.), Extreme Speech and Democracy. Oxford University Press.
[11] Burton, J. (2008). 'A Section Too Far?' Index on Censorship, 37, 115-119.
[12] Hardy, K. (2017). 'Hard and Soft Power Approaches to Countering Online Extremism'. In M. Conway, L. Jarvis, O. Lehane, S. Macdonald & L. Nouri (eds.), Terrorists' Use of the Internet: Assessment and Response. IOS Press.
[13] Pearson, E. (2018). 'Online as the New Frontline: Affect, Gender, and ISIS-Take-Down on Social Media'. Studies in Conflict & Terrorism, 41, 850-874.
[14] Macdonald, S., Correia, S. & Watkin, A. (2019). 'Regulating Terrorist Content on Social Media: Automation and the Rule of Law'. International Journal of Law in Context, 15, doi:10.1017/S1744552319000119.
[15] Citron, D. K. (2018). 'Extremist Speech, Compelled Conformity, and Censorship Creep'. Notre Dame Law Review, 93, 1050.
[16] Hazel Blears MP, Hansard, HC Debates, 9 November 2005, col 430.
[17] Forcese, C. & Roach, K. (2015). 'Criminalizing Terrorist Babble: Canada's Dubious New Terrorist Speech Crime'. Alberta Law Review, 53, 35-84.
[18] Macdonald, S. & Lorenzo-Dus, N. (forthcoming). 'Purposive and Performative Persuasion: The Linguistic Basis for Criminalising the (Direct and Indirect) Encouragement of Terrorism'.
[19] Searle, J. (1975). 'Indirect speech acts' In P. Cole & J. L. Morgan (eds.), Syntax and Semantics, Volume 3: Speech Acts. Academic Press.
[20] Terkourafi, M. (2011). 'The puzzle of indirect speech'. Journal of Pragmatics, 43, 2861-2865.
[21] van der Vegt, I., Gill, P., Macdonald, S. & Kleinberg, B. (2019). Shedding Light on Terrorist and Extremist Content Removal. RUSI.
[22] Carlile, A. (2007). The Definition of Terrorism. Cm 7052, Home Office.
[23] Anderson, D. (2014). The Terrorism Acts in 2013: Report of the Independent Reviewer on the Operation of the Terrorism Act 2000 and Part 1 of the Terrorism Act 2006. The Stationery Office.
[24] It is noteworthy, therefore, that Facebook's definition of terrorism has been described as overly broad by the United Nations Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism (Ní Aoláin 2018).
[25] Searle, J. (1975). 'Indirect speech acts' In P. Cole & J. L. Morgan (eds.), Syntax and Semantics, Volume 3: Speech Acts. Academic Press.

[26] R v Gul [2013] UKSC 64.
[27] Per Lords Neuberger and Judge (para 36).
[28] Bickert, M. & Fishman, B. (2017). 'Hard Questions: How We Counter Terrorism'. Facebook News, 15 June 2017 https://newsroom.fb.com/news/2017/06/how-we-counter-terrorism/
[29] van der Vegt, I., Gill, P., Macdonald, S. & Kleinberg, B. (2019). Shedding Light on Terrorist and Extremist Content Removal. RUSI.
[30] Grinnell, D., Macdonald, S. & Mair, D. (2017). The Response of, and on, Twitter to the Release of Dabiq Issue 15. Paper presented at the 1st European Counter Terrorism Centre (ECTC) conference on online terrorist propaganda, 10-11 April 2017, at Europol Headquarters, The Hague. https://www.europol.europa.eu/publications-documents/response-of-and-twitter-to-release-of-dabiq-issue-15
[31] Grinnell, D., Macdonald, S., Mair, D. & Lorenzo-Dus, N. (2018). Who Disseminates Rumiyah? Examining the Relative Influence of Sympathiser and Non-Sympathiser Twitter Users. Paper presented at the 2nd European Counter Terrorism Centre (ECTC) conference on online terrorist propaganda, 17-18 April 2018, at Europol Headquarters, The Hague. https://www.europol.europa.eu/publications-documents/who-disseminates-rumiyah-examining-relative-influence-of-sympathiser-and-non-sympathiser-twitter-users
[32] Note also the possibility of a separate prohibition on the sharing of terrorism-promoting content without any context or qualification. This is already Facebook's current practice (Rosen, 2019).
[33] See, e.g., Elfbrandt v Russell 384 U.S. 11 (1966) and Noto v United States 367 U.S. 290 (1961).