ACCESS NOW POSITION PAPER: A DIGITAL RIGHTS APPROACH TO PROPOSALS FOR PREVENTING OR COUNTERING VIOLENT EXTREMISM ONLINE

accessnow

Access Now is an international organisation that defends and extends the digital rights of users at risk around the world. By combining innovative policy, global advocacy, and direct technical support, we fight for open and secure communications for all.

ACKNOWLEDGEMENTS

We would like to acknowledge the considerable work that has been initiated on this subject by international human rights law experts, fellow civil society organisations and academics, companies, and government policy makers to identify a rights-respecting path on a complex policy issue. Additionally, we would like to thank the many organisations and individuals who have contributed their inputs, assistance, and feedback to this paper, including:

Courtney Radsch, Chinmayi Arun, Daphne Keller, David Sullivan, Faiza Patel, Ingvil Andersen, Jillian York, Kristin McCarthy, and Nancy Payne. This paper does not necessarily reflect their views.

The primary author of this paper is Raman Jit Singh Chima, with assistance from Amie Stepanovich, Brett Solomon, Donna Wentworth, Drew Mitnick, Estelle Massé, Javier Pallero, Peter Micek, and Lucie Krahulcova. Jamie Tomasello's insights and guidance were valuable in the conceptualisation and framing of this paper.

TABLE OF CONTENTS

I. EXECUTIVE SUMMARY

Governments, policymakers, and law enforcement across the world are showing increased interest in pushing for proactive monitoring, surveilling, censoring, or otherwise modifying certain types of online content, under the broad rubric of "preventing" or "countering" violent extremism (PVE or CVE).

These proposals risk targeting satire, journalism, activism and organising, political protest, and other forms of speech, and undercutting existing rule of law and human rights safeguards. Troublingly, several proposals have suggested that companies and web platforms should proactively modify online content and communications or create opaque and unaccountable channels of cooperation — despite the clear indication that such practices would undermine fundamental rights, the rule of law, and wider trust in the internet.[1]

Human rights around the world depends on access to a free and open internet. Policymakers, when responding to legitimate concerns about terrorism and violent extremism, must not propose and enforce policies that violate or put at risk these fundamental rights. Regardless of the rationale for a particular CVE programme, the concept of countering "violent extremism" and "extremism" online should not be used as the basis for restricting freedom of expression, nor violating the right to privacy.

Any initiative to counter violent extremism must be grounded in a definition of that term that focuses on specific criminality, avoiding sweeping generalisations with identifiers such as ethnic origin, political affiliation, etc.

The definition must be anchored in an accountable and independent legal system with adequate oversight in order to prevent abuse and ensure the right to appeal. Further, initiatives that employ tactics tantamount to surveillance must be conducted using the same human rights safeguards applicable to all communications surveillance.[2]

[1] Examples of this include the "code of conduct" discussion regarding alleged online hate speech content at the European Commission and concern about the impact on digital rights, and proposals mooted by policymakers in other countries, including the United States. See *EDRi and Access Now withdraw from EU Commission discussions,* May 31 2016, https://www.accessnow.org/edri-access-now-withdraw-eu-commission-forum-discussions/, and Reuters, *White House Lobbies Tech Leaders in War Against Online Militants*, Jan 8 2016, http://fortune.com/2016/01/08/white-house-lobbies-tech-leaders-in-war-against-online-militants/.
[2] See International Principles on the Application of Human Rights to Commuications Surveillance, May 2014, https://necessaryandproportionate.org/principles (herinafter referred to as the "Necessary and Proportionate" principles).

In order to address the lack of clarity and rights-invasive activities conducted in the area of CVE, Access Now has set out three high-level principles and subject-specific recommendations to protect users' rights:

- **Principle One: Foster dialogue and education transparently, without bias.** Efforts to counter violent extremism by promoting open dialogue or education online must be transparent and not privilege certain forms of speech.

- **Principle Two: Respect users' privacy.** Any approach for countering violent extremism that constitutes surveillance — such as social media monitoring, algorithmic content reporting, or content referral programmes — must be subject to the same normative and legal restrictions applicable to communications surveillance in other contexts.

- **Principle Three: Avoid coercion of private industry to undermine free expression protections.** Governments must not compel companies to conduct programmes to counter violent extremism, either by advancing new legislation or by threatening to screen or censor speech outside of legal process.

Our policy recommendations, in Section III, are rooted in well-established human rights law and specifically tailored for public officials and policymakers, companies, and civil society. While we do not seek with these recommendations to provide a complete guide for when actions taken for CVE are appropriate, they do provide a baseline for ensuring that human rights are not undermined in their pursuit.

II. DISCUSSION FRAMEWORK FOR CVE ONLINE AND HUMAN RIGHTS

In July 2016 the UN Human Rights Council declared that "the same rights that people have offline must also be protected online".[3] But what does this mean when we consider proposals to counter violent extremism online? Following are two key pillars for framing a human rights analysis of CVE programmes:

→ Concerns about "violent extremism" must not undermine our fundamental rights

We cannot toss out all that we have learned about protecting free expression or limiting surveillance because activity once conducted in the physical environment now takes place on the internet, or because conversations that were once hidden are now visible online.

We must work within established international human rights frameworks to evaluate any proposal for countering violent extremism that contemplates (1) interfering with online content and free expression, or (2) conducting any form of online "monitoring" or surveillance.

Any restriction on the freedom of expression must pass the established test, as detailed in the UN Human Rights Committee's General Comment 34 on the International Covenant on Civil and Political Rights, that the restrictions are "provided by law"; imposed on one of the grounds set out in Article 19(3); and conform to the strict tests of necessity and proportionality.[4] If a measure or practice that constitutes surveillance implicates "Protected Information" — that is, information that includes, reflects, arises from, or is about a person's communications, and that is not readily available and easily accessible to the general public — it must adhere to international human rights law and comparative global standards. This includes the International Principles on the Application of Human Rights to Communications Surveillance (the "Necessary and Proportionate" principles).[5]

→ Avoid sweeping definitions and unclear language

Language is powerful, and in the context of proposals to counter violent extremism online, lack of clarity can be dangerous. Vague or overbroad definitions of terms like "extremism" or "violent extremism" could easily build the foundation for human rights violations and put vulnerable communities at risk.

[3] "Affirms that the same rights that people have offline must also be protected online, in particular freedom of expression, which is applicable regardless of frontiers and through any media of one's choice, in accordance with articles 19 of the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights". UN Human Rights Council, *The promotion, protection and enjoyment of human rights on the Internet*, 27 June 2016 A/HRC/32/L.20.
[4] UN Human Rights Committee, *General Comment No. 34: Article 19 - Freedoms of opinion and expression*, 12 Sep 2011, CCPR/C/GC/34 http://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf. Such standards also exist in regional human rights systems. See e.g. - Council of the European Union, EU Human Rights Guidelines on *Freedom of Expression Online and Offline*, http://eeas.europa.eu/delegations/documents/eu_human_rights_guidelines_on_freedom_of_expression_online_and_offline_en.pdf
[5] International Principles on the Application of Human Rights to Commuications Surveillance, May 2014, https://necessaryandproportionate.org/principles (hereinafter referred to as the "Necessary and Proportionate" principles; signed by more than 200 organisations and 275,000 individuals globally).

Definitions of the term "violent extremism" vary greatly, subject to a variety of factors.[6] This term is often purported to apply to any race, religion, or ideology. In practice, however, it often ends up being applied disproportionately to people belonging to certain communities or ethnicities, a trend that special rapporteurs from the UN, IACHR, ACPHR, and OSCE noted with concern in the May 2016 Joint Declaration on Freedom of Expression and Countering Violent Extremism.[7]

Law enforcement in the UK has also expressed concern about the lack of harmonised definition for "extremism", and questioned whether anti-radicalisation efforts based on such underpinnings could be enforced.[8] More recently, the UK Parliament's Joint Committee on Human Rights noted that "[i]t is far from clear that there is an accepted definition of what constitutes extremism, let alone what legal powers there should be, if any, to combat it".[9]

If we are to understand, much less operationalise, the term "violent extremism", its definition must be narrow and focus on specific criminality, avoiding any sweeping generalisations with identifiers such as ethnic origin, political affiliation, etc. The definition must be anchored in an accountable and independent legal system with adequate oversight in order to prevent abuse and ensure the right to appeal. This requirement for anchoring the term within a legal system would also be necessary for any process to hold an organisation liable for providing material support for terrorism or to rule that it is a "designated terrorist group".[10]

Ultimately, the right to hold thoughts and views — even those considered extremist by some — must be protected unless an expression constitutes incitement to violence or falls under other specific exceptions to free speech protection under international human rights law. As UK MPs have stated in the context of parliamentary inquiry into countering violent extremism online, "the aim should be to tackle extremism that leads to violence, not to suppress views with which the Government disagrees".[11]

[6] As noted by the UN Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism: "Despite the numerous initiatives to prevent or counter violent extremism, there is no generally accepted definition of violent extremism, which remains an 'elusive concept'". *Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, 22 February 2016, A/HRC/31/65/.*
[7] Joint Declaration on Freedom of Expression and countering violent extremism, 4 May 2016, http://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=19915&LangID=E (hereinafter referred to as "Joint Declaration on Free Expression and CVE").
[8] The Guardian, *Anti-radicalisation chief says ministers' plans risk creating 'thought police'*, 24 May 2016, http://www.theguardian.com/uk-news/2016/may/24/anti-radicalisation-chief-says-ministers-plans-risk-creating-thought-police.
[9] House of Lords & House of Commons - Joint Committee on Human Rights, *Counter-Extremism: Second Report of Session 2016-17*, 20 July 2016, https://www.publications.parliament.uk/pa/jt201617/jtselect/jtrights/105/105.pdf, at 4.
[10] We note that governments and private firms are increasingly incorporating UN Security Council Counterrorism Committee designations for terrorist organisations in their policies and legal notifications, in addition to national listings.
[11] *Supra* note 9, at 3.

Indeed, all stakeholders must heed the words of caution offered by the UN Human Rights Committee in 2011 when it issued General Comment 34 and spoke of the responsibilities of ICCPR signatory states with respect to counter-terrorism measures:

> "Such offences as 'encouragement of terrorism' and 'extremist activity' as well as offences of 'praising', 'glorifying', or 'justifying' terrorism, should be clearly defined to ensure that they do not lead to unnecessary or disproportionate interference with freedom of expression. Excessive restrictions on access to information must also be avoided. The media plays a crucial role in informing the public about acts of terrorism and its capacity to operate should not be unduly restricted."[12]

Accordingly, in this paper, we use the terms "violent extremism" and "countering violent extremism" only to provide guidance and principles for avoiding violating human rights; we do not indicate agreement with the problematic framing or approach that use of these terms represents. We strongly endorse the recommendation made in the Joint Declaration on Freedom of Expression and Countering Violent Extremism that:

> "the concept of 'violent extremism' and 'extremism' should not be used as the basis of restricting freedom of expression unless they are defined clearly and appropriately narrowly".[13]

[12] *Supra* note 4, at para 46.
[13] Joint Declaration on Free Expression and CVE, *supra* note 7.

III. POLICY PRINCIPLES AND RECOMMENDATIONS

Access Now has developed three high-level policy principles to help guide stakeholders — public officials and policymakers, companies, and civil society — seeking to navigate the complex debate on countering violent extremism online, supplemented with additional specific recommendations based on these principles. Below is a list of the principles, followed by a brief discussion of each principle and specific policy recommendations for stakeholders.

- → **Principle One: Foster dialogue and education transparently, without bias.** Efforts to counter violent extremism by promoting open dialogue or education online must be transparent and not privilege certain forms of speech.

- → **Principle Two: Respect users' privacy.** Any approach for countering violent extremism that constitutes surveillance — such as social media monitoring, algorithmic content reporting, or content referral programmes — must be subject to the same normative and legal restrictions applicable to communications surveillance in other contexts.

- → **Principle Three: Avoid coercion of private industry to undermine free expression protections.** Governments must not compel companies to conduct programmes to counter violent extremism, either by advancing new legislation or by threatening to screen or censor speech outside of legal process.

PRINCIPLE ONE

Foster dialogue and education transparently, without bias. Efforts to counter violent extremism by promoting open dialogue or education online must be transparent and not privilege certain forms of speech.

Independent reporting and open, free discussion online can help defeat arguments used by those seeking to further "violent extremism".[14] It is possible that online platforms could help facilitate such open dialogue, through technical means such as enabling replies in online forums. However, more research and public discussion is needed to determine whether such an approach would be effective, as well as to ensure that any such efforts are transparent to the users. Otherwise, trust in online communications is eroded, and we risk feeding, not discouraging, extremism.

Another issue to consider is that media actors and academics working to promote "counter narratives" and open dialogue are often attacked from multiple sides in the CVE debate; some government actors treat them with suspicion or even prosecute them, while the "violent extremists" may threaten them.[15]

[14] A point also emphasised by the free expression special rapporteurs to the UN, IACHR, ACPHR, and OSCE. *See* Joint Declaration on Free Expression and CVE, *supra* note 7.
[15] These concerns are not academic. For example, the UN special rapporteur on combating terrorism noted in a recent report: "Following one case in which an individual was convicted of providing material support for Al-Qaeda by translating and posting on the Internet recruitment videos and other documents, critics decried that "ordinary people — including writers and journalists, academic researchers, translators, and even ordinary web surfers —[can] be prosecuted for researching or translating controversial and unpopular ideas". *Supra* note 6, at 14. *See also* Joint Declaration on Free Expression and CVE, supra note 7 ("Reaffirming the critical role that freedom of expression can play in promoting equality and in combating intolerance, and the essential role that the media and the Internet and other digital technologies play in keeping society informed, and stressing that limiting the space for freedom of expression and restricting civic space advances the goals of those promoting, threatening and using terrorism and violence").

While it is preferable to focus on seeking to foster more speech rather than censoring or otherwise curtailing free expression online,[16] we must keep in mind that there is no consensus as to the efficacy of any kind of traditionally understood, government-led "counter-speech" programmes, particularly given the fact that they could become delivery vehicles for propaganda.

Further, efforts operating under a CVE banner are often ineffective because they focus on national security terminology, rather than on seeking to ensure conflicts are mitigated and community trust heightened.[17]

Programmes or proposals where government agencies compel the promotion of particular CVE messages (created by governments or their partners) through online platforms and social media services, or which require the content delivered by web services to be algorithmically modified, are deeply troubling. They should not be advanced.[18]

Instead of pushing certain messages or prioritising particular content, we suggest promoting or enabling diverse voices and channels of communication online, without giving preferential treatment to particular perspectives. Strategies to accomplish this could include government grants to help non-profits that foster open dialogue get online, to provide social media training, to support non-profit advertising programmes, etc.

Transparency is key for maintaining the trust in the open communications that enable free expression and debate in our communities. Transparency must therefore define any corporate, government-run, or state-supported programmes established to combat violent extremism online, including programmes to remove content, deactivate accounts, or promote counter-narratives. If any counter-narrative messaging is paid for or otherwise supported by governments, it must be clearly labelled and attributed.

Transparency is also important to deepen our understanding of efforts to counter violent extremism. Some institutions engaging in pilot work with online platforms have recently released research and data, and so have some tech firms.[19] But more such efforts are necessary, as are institutional processes to evaluate impact.

[16] As the joint declaration on countering violent extremism from international special rapporteurs on free expression noted: "Governments should counter ideas they disagree with, but should not seek to prevent non-violent ideas and opinions from being discussed". *Supra* note 7.
[17] *See e.g.,* Dana Hadra, *Brookings Institution: A how-to on countering violent extremism,* 21 March 2016, https://www.brookings.edu/2016/03/21/a-how-to-on-countering-violent-extremism/ ("CVE efforts… must focus less on 'defeating and destroying' and more on conflict prevention and mitigation", "Policymakers… should think carefully about what they label CVE to avoid further destabilising already vulnerable communities"), *and* Access Now & Orgs., *NGO Coalition letter to White House re: Federal Support for Countering Violent Extremism Programs*, 22 April 2016, https://www.aclu.org/letter/coalition-letter-white-house-re-federal-support-countering-violent-extremism-programs ("… appropriate strategies would treat communities holistically and address a range of needs and social problems, rather than through the singular lens of national security or law enforcement. In any event, government programs and partnerships cannot target a particular religious community or determine participants by reference to religion and/or national origin. They may not advance a particular set of religious beliefs while suppressing others").
[18] Article 20.1 of the ICCPR clearly states that "Any propaganda for war shall be prohibited by law". Mechanisms to help disseminate government propaganda in the channels of internet platform services and online content - even for peace - could go down a path which risks violating Article 20, particularly when available as a tool to less scrupulous governments.
[19] *See e.g.* Institute for Strategic Dialogue, *The Impact of Counter-Narratives, Insights from a year-long cross-platform pilot study of counter-narrative curation, targeting, evaluation and impact*, July 2016, http://www.strategicdialogue.org/wp-content/uploads/2016/08/Impact-of-Counter-Narratives_ONLINE.pdf, and Twitter Policy Blog, An update on our efforts to combat violent extremism, 18 Aug 2016, https://blog.twitter.com/2016/an-update-on-our-efforts-to-combat-violent-extremism.

Finally, in evaluating proposals for countering violent extremism online, companies should be mindful of their responsibilities under the UN Guiding Principles on Business and Human Rights.[20] That entails consulting affected stakeholder groups and examining the impact their platforms and services may have on fundamental rights via CVE or other programmes, and taking action to prevent or mitigate any adverse impacts directly related to their products. States, for their part, should not impose liability strictures upon companies that encourage them to stay ignorant of possible human rights infringements.

Specific stakeholder recommendations:

For policymakers and public officials:

1. Adopt policy frameworks and legislative measures that favour internet-enabled independent journalism, blogging platforms, and investigative reporting; review existing legal measures and prosecution policies to prevent clamping down on this critical channel for disseminating facts and supporting dialogue.

2. Support — and prevent the chilling of — efforts to drive forward genuine academic inquiry conducted via the internet on issues connected with "violent extremism".

3. Explore ways to support efforts to create further dialogue using the internet, without preferential treatment for how content is disseminated. This could include methods such as helping genuine dialogue-supporting organisations and community leaders establish an online presence, funding public advertising (for example, providing publicly disclosed advertising grants to nonprofits or independent institutions that promote inter-community dialogue), or developing additional outreach and communications channels.

For companies:

1. Ensure that any efforts to provide support to groups working to counter violent extremism are transparent, sound in methodology, and do not endanger the furtherance of human rights. There is an urgent need for more transparency and understanding of impact for company programmes or pilot efforts in this regard.

2. Undertake further research and dialogue to explore how product design efforts — such as enabling direct replies in online platforms — can support meaningful dialogue, discourage echo chambers, and reduce speech that directly incites violence.

For public-private partnerships:

1. Government and private sector partnerships for countering "violent extremism" should at a minimum follow transparency and disclosure norms in this space — including following regulations for national lobbying or state propaganda. There should be ongoing commitment to oversight of any such partnerships by independent government oversight agencies, civil society organisations, human rights experts, national human rights institutions, and multi-stakeholder groups. Any counter-narrative messaging paid for or otherwise supported by governments must be clearly labelled and attributed.

[20] United Nations - Office of the High Commissioner on Human Rights, *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect, and Remedy" Framework*, 16 June 2011, http://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf.

PRINCIPLE TWO

Respect users' privacy. Any approach for countering violent extremism that constitutes surveillance — such as social media monitoring, algorithmic content reporting, or content referral programmes — must be subject to the same normative and legal restrictions applicable to communications surveillance in other contexts.

Many online CVE-related policy proposals entail social media monitoring, algorithmic content reporting, or content referral programmes, often in order to identify content that may trigger content or account removal, counter-narrative mechanisms, or official legal investigations. Surveillance of this sort can have a disparate impact on users at risk, including but not limited to vulnerable groups such as journalists and activists, communities of colour, persecuted religious groups, and members of LGBTQI communities. Human rights experts have specifically noted the concern raised by basing surveillance on ethnic or religious profiling, and the targeting of whole communities rather than specific individuals.[21]

If governments deputise companies and individual users to conduct monitoring — or undertake the monitoring themselves — and the surveillance implicates Protected Information, there must be adherence to international human rights law and comparative global standards, including the International Principles on the Application of Human Rights to Communications Surveillance (the "Necessary and Proportionate" principles), which has 13 principles: Legality, Legitimate Aim, Necessity, Adequacy, Proportionality, Competent Judicial Authority, Due Process, User Notification, Transparency, and Public Oversight, Integrity of Communications and Systems, Safeguards for International Cooperation, and Safeguards Against Illegitimate Access.

Government-run or state-supported programmes to track and monitor Protected Information can have serious repercussions. Even publicly available information can become Protected Information when the monitoring is pervasive, under the terms of the "Necessary and Proportionate" principles. To monitor social media en masse is to treat all users like suspects, which has a chilling effect on human rights such as the rights to privacy, free speech, and access to information. It also discourages trust in the internet economy.

In practice, such large-scale monitoring of a vaguely defined category of content related to "violent extremism" can — and often is — applied with a discriminatory impact that adversely affects people in social movements, such as those advocating for racial and gender equality and criminal justice.

Tools such as algorithmic content flagging also carry high risks with respect to the likelihood of false positives.[22] This may further exacerbate the negative impact of such programmes, including further radicalising communities, silencing others, and undermining global trust in the opportunities for communication and open dialogue that the internet provides.

Specific stakeholder recommendations:

For policymakers and public officials:

1. Governments must not force or request online platforms to undertake actions regarding user data disclosure or other surveillance measures that are outside of rule-of-law processes that comply with international human rights law and policy (including the "Necessary and Proportionate" principles).[23]

[21] Joint Declaration on Free Expression and CVE, *supra* note 7.
[22] See further discussion of this later in this paper under the issue recommendations under policy principle 3, i.e. "Governments must not force or request online platforms to undertake actions regarding user data disclosure or other surveillance measures that are outside of rule-of-law processes that comply with international human rights law and policy (including the 'Necessary and Proportionate' principles)".
[23] The "Necessary and Proportionate" Principles show how existing human rights law applies to digital surveillance. They have been widely adopted and signed by more than 200 organisations and 275,000 individuals globally including legal experts, political parties, and elected officials.

2. Given the potential for social media monitoring to interfere with the rights to free expression and privacy when it pertains to Protected Information, any such practices must be provided for by law in a manner compatible with the "Necessary and Proportionate" principles. In particular, this includes — but is not limited to — the following:

   2.1. When information that is collected is Protected Information, then it should only be collected when it is both necessary and proportionate to a legitimate aim to do so. State action to authorise the collection of Protected Information must respect the "Necessary and Proportionate" principles, with a stepwise process regarding government application for information, judicial consideration, search, appeals and remedies, and international cooperation (if applicable).

   2.2. Government agents should never seek access to Protected Information outside of legal process, and particularly not through misleading methods, such as by creating fake profiles to follow or "friend" a user.

PRINCIPLE THREE

Avoid coercion of private industry to undermine free expression protections. Governments must not compel companies to conduct programmes to counter violent extremism, either by advancing new legislation or by threatening to screen or censor speech outside of legal process.

CVE-related proposals that include plans for proactive removal of content, manual or algorithmic "de-prioritisation" of content, or other types of interference with content, may appeal to governments concerned about violent extremism. However, these approaches directly impact the right to free expression. And just like there is no "magic key" to ensure that only a trusted government can break encryption to access Protected Information, there is no "magic eraser" to allow companies automatically to identify and remove or de-prioritise *only* illegal content.

Governments may also pursue mass take-down requests for content that is alleged to encourage violent extremism. This includes the increasingly popular practice of creating so-called internet referral units, through which a large number of takedown requests are sent to companies outside the channel for legal removal requests.[24] Such mass take-downs can often be counterproductive, risking silencing voices seeking to respond to or counter violent extremist narratives. Content should not be removed until it is specifically adjudicated as being illegal, in line with international standards in this area, including General Comment 34. Mass take-down initiatives that take place outside of legal process frustrate corporate transparency and are not likely to deter the cultivation of "violent extremism", and in fact may encourage it, inflaming resistance and helping "violent extremist" recruiters discredit platforms that might otherwise support online expression and debate.[25]

[24] For examples of this and the concerns triggered for digital rights, see Access Now, *Europol's Internet Referral Unit risks harming rights and feeding extremism*, 17 June 2016, https://www.accessnow.org/europols-internet-referral-unit-risks-harming-rights-isolating-extremists/.

[25] *See e.g.*, Kate Ferguson, Partnership for Conflict, Crime and Security Research University of East Anglia, *Countering violent extremism through media and communication strategies: A review of the evidence*, 1 March 2016, http://www.paccsresearch.org.uk/wp-content/uploads/2016/03/Countering-Violent-Extremism-Through-Media-and-Communication-Strategies-.pdf ("VE propaganda online has expanded in the face of CVE takedowns and counternarrative strategies"), *and* Colin Baulke, Mackenzie Institute, *The Nature of the Platform: Dealing with Extremist Voices in the Digital Age*, 8 May 2016, http://mackenzieinstitute.com/nature-platform-dealing-extremist-voices-digital-age/ ("... and to further complicate the problem, the impact of successful takedown campaigns is murky. In some extremist online circles, including ISIS, users view having a suspended account as a badge of honour. Essentially, increased suspensions equate to greater legitimacy.")

If a company engages in a CVE programme, the company and those who review content (whether employees or contractors) cannot be tasked with the primary duty of evaluating the legality of content in the absence of rule-of-law mechanisms. When companies review complaints regarding content, it's necessary for staff to be well-trained to consider context and other factors. If a company uses content-flagging tools for a CVE programme, use of these tools should be limited to drawing reviewers' attention to content, not automatically flagging and taking down content, nor automatically suspending accounts. These reviewers must receive training on applying human rights standards — within the framework of local contexts — in addition to other kinds of support and resources.

Additionally, reviewers cannot be placed in situations where they are asked to act as editors, choosing to keep some categories of content online while removing others based on "countering violent extremism" strategies.[26] Such practices can result in reviewers or moderators knowingly or unknowingly chilling free expression, as well as suppressing satire or other kinds of speech seeking to respond to or counter calls to violent extremist action. Their role should remain focused on taking down content when they are notified that it explicitly violates their terms of service, or when they receive legal process requiring access to content be suspended or disabled.

It's misleading to argue for countering violent extremism online using technical solutions such as filtering or proactive content takedown simply because they're used in other situations (for example, in the context of removing child sexual abuse material). These methods are also a poor policy choice. They have a demonstrably high false-positive rate (particularly for content outside of specifically blacklisted child sexual material), and do not suit situations that lack a clear definition for content, context, or legal mandate.[27]

Even in "emergency" situations, we cannot suspend human rights protections. Governments and public officials are sometimes confronted with situations pertaining to online content and violent extremism that they regard as fast moving, and with potential negative consequences for the safety of citizens and public order. Policy planning for such situations should be underpinned in legal mechanisms that allow for rapid responses while ensuring that procedural safeguards are in place and the requirements of international human rights law are met.[28] It is not acceptable to implement state-operated mechanisms or other arrangements in the absence of law. General Comment 34 on the ICCPR notes that the

[26] Such steps must also be avoided due the impact they would have on the legal position on internet intermediaries, given that many jurisdictions across the world possess legal provisions that provide a limited "safe harbour" protection to intermediaries for third party or user generated content — but often subject to the requirement that they do not interfere or editorially engage with the content in question.
[27] Many of these proposals also fail to note that the usage of such technical tools to detect and report child sexual abuse material was developed in the specific context of legal regimes across most countries criminalising the very possession of such material, under provisions meant to combat child pornography or sexual abuse.
[28] The framing of any such legal models must be approached cautiously. Many proposals may grant certain public officials the power to issue emergency web content blocking orders, which are then post-facto reviewed by review committees or other authorities. One example of this is Section 69A of the Indian Information Technology Act and its implementing rules, which allow the issuance of emergency blocking orders which have to be later examined by a review committee. The operation of this review committee and the emergency blocking process has been criticised for being opaque and limiting itself to procedural review without any examination of the validity of blocking requests. *See* Human Rights Watch, *Stifling Dissent The Criminalization of Peaceful Expression in India*, 24 May 2016, https://www.hrw.org/report/2016/05/24/stifling-dissent/criminalization-peaceful-expression-india.

right to free expression cannot be derogated from even during a public emergency,[29] and specifically requires that when considering restrictions on free speech:

> "the restrictions must be 'provided by law';
>
> they may only be imposed for one of the grounds set out in subparagraphs (a) and (b) of paragraph 3 [of Article 19]; and
>
> they must conform to the strict tests of necessity and proportionality.
>
> Restrictions are not allowed on grounds not specified in paragraph 3 [of Article 19], even if such grounds would justify restrictions to other rights protected in the Covenant. Restrictions must be applied only for those purposes for which they were prescribed and must be directly related to the specific need on which they are predicated... "[30]

The special rapporteurs on free expression to the UN, OSCE, OAS, and ACHPR have built on the guidance of General Comment 34 in their Joint Declaration on Free Expression and Countering Violent Extremism, in which they state:

> "c) Any restrictions on freedom of expression should comply with the standards for such restrictions recognised under international human rights law. In compliance with those standards, States must set out clearly in validly enacted law any
>
> restrictions on expression and demonstrate that such restrictions are necessary and proportionate to protect a legitimate interest.
>
> d) Restrictions on freedom of expression must also respect the prohibition of discrimination, both on their face and in their application.
>
> e) Restrictions on freedom of expression must be subject to independent judicial oversight."

It follows that we must also reject government efforts seeking to hold liable companies running online platforms, social media, or communications services on grounds such as incitement, defamation of religion, or material support to violent "extremists".[31] Indeed, the special rapporteurs' joint declaration stated that they are:

> "*Concerned* about pressure on private companies, and especially social media networks, to 'cooperate' in reporting on those whom they suspect of radicalisation

[29] "Furthermore, although freedom of opinion is not listed among those rights that may not be derogated from pursuant to the provisions of article 4 of the Covenant, it is recalled that, 'in those provisions of the Covenant that are not listed in article 4, paragraph 2, there are elements that in the Committee's opinion cannot be made subject to lawful derogation under article 4'. Freedom of opinion is one such element, since it can never become necessary to derogate from it during a state of emergency". *Supra* note 4, at para 5.
[30] *Supra* note 4, at para 22.
[31] Courts have also begun indicating their refusal to accept broad arguments calling for online platforms and social media services to be held liable for "providing material support" to terrorist organisations simply because extremist organisations sign up and use their services. *See e.g.* Bloomberg, *Twitter Ruled Not Liable for ISIS Tweets Leading to Attack*, 11 Aug 2016, http://www.bloomberg.com/news/articles/2016-08-10/twitter-ruled-not-liable-for-isis-tweets-leading-to-attack, reporting on the ruling by the US District Court for the Northern District of California in *Tamara Fields v. Twitter Inc.*, Case No. 16-cv-00213-WHO (N.D. Cal. Aug 10, 2016), https://casetext.com/case/fields-v-twitter-inc (dismissing a complaint brought against Twitter arguing that it should be held liable broadly for material support to the Islamic State of Iraq and Syria for a shooting at a law enforcement training centre in Amman which resulted in the deaths of US government contractors).

and the fact that CVE/PVE is increasingly being used by companies to justify measures restricting content, sometimes without being transparent or consistent about the rules and the kinds of expression that are being limited."[32]

For any corporate policy that regards content removal, transparency is imperative — particularly surrounding the specific reasons for removing the content. CVE efforts also demand additional oversight and mechanisms for redress. For example, when a company takes down a user's content or suspends the user's account (temporarily or permanently), it should inform the user of the rationale and provide information about any processes for appealing the decision. This applies regardless of whether the action is self-initiated or prompted by a third party; affected parties must be notified, and the company must report in the aggregate, when content is removed or access is restricted. Users must be provided meaningful access to remedy through an appeals mechanism.[33] Additionally, companies must ensure that when governments request that content be removed, whether because it violates the terms of service or is illegal, those requests are included in transparency reports and categorised as such.

Specific stakeholder recommendations:

For policymakers and public officials:

1. Mass take-downs of content are often counterproductive, and should not be implemented until content is specifically identified as unlawful or illegal. Such efforts do not deter the cultivation of "violent extremism", and in fact may encourage it, inflaming resistance and helping "violent extremist" recruiters discredit platforms that might otherwise support online expression and debate.

2. Governments must not expand or influence Terms of Service agreements in ways that "deputise" or pressure corporations to carry out the aims of the state.

3. Government use of platform "flagging" tools or other automated processes should not be allowed to become a channel to bypass corporate transparency or rule of law processes and human rights safeguards that normally govern governmental powers regarding restricting speech and expression.

4. Government steps targeting the removal of violent extremist content must operate within the restrictions placed by human rights standards and fundamental rights,[34] particularly with respect to the following:

   4.1. Governments must not force or request platforms to remove content unless

      a. it has been adjudicated to be unlawful or specifically ordered to be removed under rights-respecting legal process; and

      b. mechanisms for notice and redress for the accused speaker is provided for, within the relevant laws.

[32] Joint Declaration on Free Expression and CVE, *supra* note 7.
[33] For trends on such requests and discussions of proposed practices, see e.g. - OnlineCensorship.org, *Unfriending Censorship: Insights from four months of crowdsourced data on social media censorship*, https://onlinecensorship.org/news-and-analysis/onlinecensorship-org-launches-first-report-download; *and* Erica Newland, Caroline Nolan, Cynthia Wong, and Jillian York, *Account Deactivation and Content Removal: Guiding Principles and Practices for Companies and Users*, https://cdt.org/files/pdfs/Report_on_Account_Deactivation_and_Content_Removal.pdf.
[34] Spelt out in detail by General Comment 34 of the UN Human Rights Committee and its subsequent global usage. *Supra* note 4.

For companies:

1. Companies should indicate the reason for removing content or banning accounts, rather than merely communicating that a decision has been implemented.

2. Reporting/flagging tools, and appeal mechanisms when content or users are flagged, must meet high standards for transparency, accountability, and human rights remedy. This also extends to programmes to deputise "super-users", whether in or outside government, to report or flag allegedly violent extremist content on online platforms. Multi-stakeholder bodies — if properly constituted with the active and meaningful engagement of civil society — could oversee the development and deployment of these tools and mechanisms.

3. Companies should be wary about deploying intrusive programmes that implement proactive filtering and reporting of content using "voluntary" models that circumvent national law and international standards for interfering with free speech and expression.

For civil society and academia:

1. Many countries already have official policies or legal regimes in place regarding online speech. Civil society and academia should be watchful of these policies and how they are interpreted and implemented, to ensure they are not used to silence speech or quell protest.

IV. CONCLUSION

Stakeholders — including public officials, policymakers, companies that run web platforms and social media services, and civil society — must be extremely cautious regarding proposals advanced under the banner of countering violent extremism online. Many of the proposals now under consideration or already deployed threaten internationally protected human rights, including the rights to privacy and freedom of expression.

Given the varying legal standards globally even for defining the terms "violent extremism" or "extremism", standards for scrutiny of CVE programmes must be high. Any CVE proposal or practice that implicates human rights must be grounded in clearly defined legal provisions, in the context of an accountable and independent legal system with adequate oversight. Partnerships between government agencies and technology companies — or the "voluntary" CVE arrangements that technology firms may consider or be pressured into accepting — cannot become loopholes to circumvent human rights scrutiny and accountability to rule-of-law institutions.

It is also imperative that CVE efforts do not become channels by which governments directly or indirectly pressure online platforms to privilege certain speech, or otherwise interfere with how people access information online. If government agencies and companies forge partnerships and conduct programmes in this area, they must reject approaches that only favour particular perspectives, and commit to increased transparency and oversight. State-directed actions that interfere with online content or otherwise impact the ability of users to freely express themselves and access information are subject to the limitations placed under human rights law, including those specified by Article 19 of the ICCPR.

We must also stand on guard against using technical solutions, such as filtering or proactive content takedowns, for CVE efforts, simply because they are already in use for other purposes; they would be risky and ill-suited to CVE programmes given the lack of clear definitions for content, context, or legal mandate.

It's not only free expression that is at risk. Some CVE proposals argue for approaches that are tantamount to surveillance, threatening the right to privacy. If a programme implicates the Protected Information of users, it must be conducted within the bounds of internationally recognised standards for oversight of surveillance. It is not exempt from human rights scrutiny merely because it is proposed under a CVE banner.

The internet can help foster education, provide access to knowledge, and open channels for dialogue that can prevent conflict and counter incitement to violence. However, some proposals for countering violent extremism online would undermine the very freedom and openness that we value, and that make the internet an empowering platform for all the world's people. If we do not protect that freedom and openness, it will destroy trust in the internet globally. That would play right into the hands of those who wish to inflame conflict — feeding, not discouraging, extremism.

For more information, please contact:

Raman Jit Singh Chima
Global Policy Director
raman@accessnow.org