GCTF
GLOBAL COUNTERTERRORISM FORUM

POLICY TOOLKIT

The GCTF Zurich-London Recommendations on Preventing and Countering Violent Extremism and Terrorism Online

POLICY TOOLKIT

The GCTF Zurich-London Recommendations on Preventing and Countering Violent Extremism and Terrorism Online

Contents

Background

Since its inception, the Internet has offered innumerable opportunities for society to facilitate communication and access to information, economic development, as well as participation in society. A free, open, secure, stable, accessible and peaceful digital environment is essential for all and requires effective cooperation among States to reduce risks to international peace and security.[1] However, violent extremist and terrorist groups and individuals use the Internet, especially social media platforms, to spread propaganda, disseminate enabling material, fundraise, intimidate, train, radicalize, recruit and incite others to commit violent extremist and terrorist acts.

The United Nations (UN) General Assembly (GA) noted the importance of cooperation among stakeholders in implementing the UN Global Counter-Terrorism Strategy, including among States, international, regional and sub-regional organizations, the private sector and civil society, to address the increasing use of information and communication technologies (ICT) by terrorists and their supporters, while respecting human rights, fundamental freedoms and complying with international law and the purposes and principles of the UN Charter.[2]

The UN GA stressed that it is essential to develop the most effective means to counter terrorist propaganda, incitement and recruitment, including through the Internet, in compliance with international law, including international human rights law. In addition, the UN GA recommended that States consider the implementation of relevant recommendations of the UN Secretary General's *Plan of Action to Prevent Violent Extremism* as applicable to the national contexts, which identified strategic communications, the Internet and social media as key action areas in preventing and countering violent extremism and terrorism.[3]

At the Seventh GCTF Ministerial Plenary Meeting in New York on 21 September 2016, a review and assessment process of existing governmental good practices and lessons learned in preventing and countering violent extremism and terrorism online was endorsed by the GCTF Members as part of the GCTF's *Initiative to Address the Life Cycle of Radicalization to Violence*. This resulted in the formal endorsement of the GCTF *Zurich-London Recommendations on Preventing and Countering Violent Extremism and Terrorism Online* (Zurich-London Recommendations) in September 2017.

This Initiative was based on the conviction that governments should take action and support appropriate actions by the ICT industry and civil society to prevent and counter the misuse of the digital space, especially the Internet and social media platforms, for violent extremist and terrorist purposes. The non-binding Zurich-London Recommendations compile a non-exhaustive list of good practices for governments for how strategic communications and social

1 Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security (A/70/174), para 2, 22 July 2015.
2 UN Global Counter-Terrorism Strategy Review (A/RES/70/291), para 42, 19 July 2016.
3 UN Global Counter-Terrorism Strategy Review, paras. 42f. and 40; UN Secretary-General's Plan of Action to Prevent Violent Extremism (A/70/674), para. 55, 24 December 2015.

media aspects can be used in preventing and countering violent extremism and terrorism online while also respecting human rights, fundamental freedoms and the principle of the rule of law.

In 2018, GCTF Members endorsed the launch of an Initiative by Australia, Switzerland and the United Kingdom that seeks to operationalize the Zurich-London Recommendations providing policy-makers and governmental experts with guidance on good governmental practices, case studies, and references to existing international and regional initiatives and practices in preventing and countering violent extremism and terrorism online.

This Policy Toolkit pursues the following objectives:

- To provide governmental experts and policy-makers with access to information on policies and current trends in preventing and countering violent extremism and terrorism online.
- To promote actions taken by governments to respect human rights and fundamental freedoms such as privacy and the freedom of expression, association, peaceful assembly, and religion or belief as well as the need to preserve a free flow of information and a free and open Internet.
- To foster efficient and sustainable collaboration between governments, ICT companies and civil society based on the principle of shared responsibility in preventing and countering violent extremism and terrorism online.
- To stimulate innovation by referring to good practices and lessons learned that may lie outside preventing and countering violent extremism and terrorism online, but that are nevertheless of relevance.

TARGET AUDIENCE

The target audience of this Policy Toolkit is GCTF Members as well as GCTF Key Partners and any other governments interested in preventing and countering violent extremism and terrorism online.

Acknowledging the shared responsibility between Governments, ICT companies and civil society in preventing and countering violent extremism and terrorism online, this Policy Toolkit is also addressed to the experts within these two latter fields.

METHODOLOGY

The Policy Toolkit seeks to provide a practical and user-friendly guide for policy-makers and experts in preventing and countering violent extremism and terrorism online. Importantly, the Policy Toolkit is non-exhaustive and intends to offer a point of reference for good practices and case studies.

The Policy Toolkit builds on the good practices identified in the Zurich-London Recommendations. Case studies as well as the examples of stakeholder practices identified therein do not promote a specific approach to preventing and countering violent extremism and terrorism online, but

were selected because they are relevant illustrations and provide guidance for what needs to be considered for successful implementation.

The Zurich-London Recommendations divide responses to violent extremism and terrorism on the Internet into two approaches:

1. Content-based responses: Government efforts to address the availability and accessibility of violent extremist and terrorist propaganda through international cooperation and to engage with private companies to counter terrorism and violent extremism online on a collaborative basis, including content reporting, removal, filtering and appropriate regulation/legislation.
2. Communications-based responses: Government efforts to support or assist in challenging the appeal of violent extremist and terrorist propaganda through strategic communications, including supporting civil society organizations to use counter- and alternative narratives both online and offline.

When firmly embedded in a whole-of-government and whole-of-society strategy to preventing and countering violent extremism and terrorism on the Internet, these two approaches can contribute to a more comprehensive approach to preventing and countering violent extremism and terrorism in general.

As such, any government-led strategy on preventing and countering violent extremism and terrorism online should set clear and measurable objectives and be underpinned by a well-defined "theory of change", articulating how and why both content and communications-based responses contribute to the objectives set out in the strategy.

Content-Based Responses:

1. Development and Adoption of Content-Related Legislation and Policies

This Chapter is intended to support law- and policy-makers as well as practitioners in the development and adoption of content-related legislation and policies to effectively prevent and counter violent extremism and terrorism online. In particular, this Chapter addresses how governments can adopt legal provisions to prevent and counter the misuse of the Internet for violent extremist and terrorist purposes, while respecting international human rights law, inter alia the freedom of expression and the right to privacy. The chapter is divided into two sub-sections: Principles and Guidelines; and Policy Design.

Relevant Good Practices from the London-Zurich Recommendations

Good Practice 1: *To adopt and implement law and policy frameworks at the national level to prevent and counter violent extremism and terrorism online.*

Good Practice 7: *To adopt laws, regulations, and policies that address the availability and accessibility of violent extremist and terrorist content on the Internet.*

Good Practice 8: *To take into account any applicable existing international standards and/or principles when addressing the availability and accessibility of violent extremist and terrorist content on the Internet and social media platforms.*

INTRODUCTION

States bear the primary responsibility in preventing and countering violent extremism and terrorism as part of a whole-of-society response. At present, only some states have adopted legal provisions that criminalize incitement to violence; substantially more states dispose of legal provisions regarding the commission of a terrorist act, glorification of terrorism, and the "apology for terrorism". The adoption or updating of legislation consistent with international human rights law to provide a legal basis to address violent extremist and terrorist content online is essential for ensuring that all relevant actors have clear obligations and effective guidelines preventing and countering violent extremism and terrorism online.

States are obliged to ensure that private actors do not breach any national or international laws in the course of their work. The development and implementation of an effective national legislative framework, be it through new content-related legislation or by updating existing texts with content-related elements, is a vital starting point for all states to ensure that unlawful content online is addressed, and that ICT companies are effectively compelled to prevent and counter violent extremism and terrorism online.[4]

A: Principles & Guidelines

International Instruments: Principles & Guidance

There are a number of international instruments which enunciate relevant standards and principles that states should consider when developing legislation to prevent and counter violent extremism. Compliance with international obligations has consistently been emphasized by international instruments. For example, the UN Global Counter-Terrorism Strategy notes that states and other relevant actors should address the increasing use of ICT by terrorists and their supporters in accordance with human rights, fundamental freedoms and complying with international law and the purposes and principles of the UN Charter.[5]

There has been consistent emphasis on the need to comply with international human rights law, and in particular, the right to freedom of expression and the right to privacy (to be discussed further below). In the *2016 Joint Declaration on Freedom of Expression and countering violent extremism*[6], the UN Special Rapporteur on Freedom of Expression, and his counterparts within the OSCE, the OAS, and the ACHPR, recommended that:

> *2. Special Recommendations (...)*
>
> *(a) States should not subject Internet intermediaries to mandatory orders to remove or otherwise restrict content except where the content is lawfully restricted*

4 For the purposes of the present tool, the term legislation will be used to refer to any law, regulation, rule, text or other instrument having force of law or a binding character in domestic contexts.
5 UN Global Counter-Terrorism Strategy Review, 2016.
6 *Joint Declaration on Freedom of Expression and Countering Violent Extremism* adopted by the UN Special Rapporteur on Freedom of Opinion and Expression, the OSCE Representative on Freedom of the Media, the OAS Special Rapporteur on Freedom of Expression and the ACHPR Special Rapporteur on Freedom of Expression and Access to Information, 03 May 2016. See also: *Joint Declaration on Challenges to the Freedom of Expression in the Next Decade*, 10 July 2019.

in accordance with the standards outlined above. States should refrain from pressuring, punishing or rewarding intermediaries with the aim of restricting lawful content (...)

(j) States should not adopt, or should revise, laws and policies which involve the following:

(i) Blanket prohibitions on encryption and anonymity, which are inherently unnecessary and disproportionate, and hence not legitimate as restrictions on freedom of expression, including as part of States' responses to terrorism and other forms of violence.

(ii) Measures that weaken available digital security tools, such as backdoors and key escrows, since these disproportionately restrict freedom of expression and privacy and render communications networks more vulnerable to attack.

Furthermore, the 2011 UN Guiding Principles on Business and Human Rights (UNGP), implement the 2008 UN "Protect, Respect and Remedy" Framework, provide additional guidance on the duties of states and responsibilities of companies to enhance their standards and practices with regard to business and human rights.[7] The first pillar of the Framework is the state duty to protect against human rights abuses committed by third parties, including business, in their territory or jurisdiction, through appropriate legislation and policies. States have the primary role in preventing and addressing corporate-related human rights abuses.

The second pillar is the corporate responsibility to respect human rights: in their actions, third parties should not infringe on the rights of others, and address adverse human rights impacts with which they are involved. While respecting rights is not an obligation that international human rights law imposes directly on third parties, it is now a common key element in nearly all voluntary and soft-law instrument related to corporate responsibility and endorsed by the Human Rights Council. Furthermore, the responsibility for business to respect rights might be reflected in national law already. Both government and corporate responsibility are addressed in the annual report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression of 6 April 2018 that addresses the regulation of user-generated online content.[8]

Finally, effective grievance mechanisms – the third pillar – play an important role in both the state duty to protect and the corporate responsibility to respect. As part of their duty to protect, states must take appropriate steps within their territory and/or jurisdiction to ensure that when such human rights abuses by business occur, those affected have access to effective remedy through judicial, administrative, legislative or other appropriate means.

7 *UN Guiding Principles on Business and Human Rights*, 2011.
8 *Annual Report of the Special Rapporteur to the Human Rights Council on online content regulation* (A/HRC/38/35), 06 April 2018.

Freedom of Expression

States should ensure that any legislation that addresses the availability and accessibility of violent extremist and terrorist content online, and conforms with international standards and principles, in particular in regards to the freedom of expression. The right to freedom of expression is a pillar of international human rights law and integral for the full enjoyment of other human rights, such as the rights to freedom of peaceful assembly and association. This universal right is enshrined in Article 19 of the Universal Declaration of Human Rights (UDHR), which states that: "Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers."

The guarantees of the freedom of expression have been further developed in Article 19 of the International Covenant on Civil and Political Rights (ICCPR)[9]. The freedom of expression has also been incorporated in a number of regional human rights law instruments such as the European Convention on Human Rights (ECHR), Inter-American Convention on Human Rights (IACHR), and the African Charter on Human and Peoples' Rights (ACHPR).

Restrictions on the Freedom of Expression

Article 19(3) ICCPR provides for two specific limitations for restricting freedom of expression: the respect of the rights or reputations of others, and the protection of national security, public order (*ordre public*), public health or morals. It is important to emphasize that the limitation clause provided by Article 19(3) must be interpreted restrictively. The *Siracusa Principles on the Limitation and Derogation of Provisions in the ICCPR* offer useful guidance on the conditions laid down by Article 19(3) to limit freedom of expression: in the context of preventing and countering violent extremism and terrorism online, safeguarding national security and/or public order are the most relevant reasons to restrict the freedom of expression. According to the Siracusa Principles, the derogations and limitations on the grounds of national security "may be invoked to justify measures limiting certain rights only when they are taken to protect the existence of the nation or its territorial integrity or political independence against force or threat of force."[10] The justification of national security "cannot be used as a pretext for imposing vague or arbitrary limitations and may only be invoked when there exists adequate safeguards and effective remedies against abuse."[11] The justification of human rights derogations and limitations by the protection of public order can be adopted only with the objective of preserving "the sum of rules which ensure the functioning of society or the set of fundamental principles on which society is founded."[12] In this regard, the Siracusa Principles emphasize that "respect for human rights is part of public order (*ordre public*)."[13]

9 Article 19(2) of the ICCPR provides the most comprehensive guarantees of the freedom of expression, the freedom to seek, receive and impart information and ideas of all kinds. Freedom of expression protects all forms of expression, including spoken, written, sign language and non-verbal expression, such as images, and the means of their dissemination, including books, newspapers, pamphlets, posters, banners, audio-visual as well as electronic and internet-based modes of expression. Please note that not all GCTF Member States have signed or ratified the ICCPR.

10 UN Commission on Human Rights, *The Siracusa Principles on the Limitation and Derogation Provisions in the International Covenant on Civil and Political Rights* (E/CN.4/1985/4), para 29, 28 September 1984.

11 *Ibid*., para 31.

12 *Ibid*., para 22.

13 *Ibid.*

The UN Human Rights Committee's General Comment no. 34 stipulates that there should be a clear definition of offences such as "encouragement of terrorism", "extremist activity" and "praising", "glorifying", or "justifying" terrorism, to ensure that there is no "unnecessary or disproportionate interference" with the freedom of expression.[14]

The UN Special Rapporteur on Freedom of Expression recommends that restrictions should only be adopted on a case-by-case basis and conform with the requirements of legality, necessity, proportionality and legitimacy[15]:

- **Legality:** Any restriction must be provided by law. Such laws must be adopted by regular legal processes and formulated with sufficient precision. In addition, such laws must be made accessible to the public and must provide sufficient guidance to those responsible for executing these laws. Any specific limitation of a right should follow due process provisions stipulated in national legislation and be overseen by independent review bodies, notably courts.
- **Necessity and proportionality:** Any restrictions must be necessary, and the least intrusive means to achieve the legitimate purpose. According to the Special Rapporteur on Freedom of Expression, social media platforms should "disclose data and examples that provide insight into the factors they assess in determining a violation, its severity and the action taken in response."[16] Furthermore, in the context of hate speech, "explaining how specific cases are resolved may help users better understand how companies approach difficult distinctions between offensive content and incitement to hatred, or how considerations such as the intent of the speaker or the likelihood of violence are assessed in online contexts."[17]
- **Legitimacy:** Any restrictions should fall within the two specific limitations in Article 19(3) of the ICCPR: the respect of the rights or reputations of others, and the protection of national security, public order (*ordre public*), public health or morals. The limitations in Article 19(3) should be interpreted restrictively. For example, derogations or limitations of freedom of expression on national security grounds should only be taken to protect a state's existence, territorial integrity or political independence against the use or threat of force.[18] National security grounds should not be used "as a pretext for imposing vague or arbitrary limitations."[19] The public order ground may only be used where the functioning of society or its fundamental principles need to be preserved.[20]

Right to Privacy

The right to privacy needs to be protected when monitoring online content to detect terrorist and violent extremist sympathizers, recruiters or terrorist plots. Monitoring of online content,

14 UN Human Rights Committee, *General comment no. 34, Article 19, Freedoms of opinion and expression* (CCPR/C/GC/34), para 46, 12 September 2011.
15 UN Human Rights Council, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression* (A/HRC/29/32), para 57, 22 May 2015.
16 UN Human Rights Council, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, 6 April 2018, A/HRC/38/35, para. 47.
17 *Ibid*., para. 47.
18 UN Commission on Human Rights, *The Siracusa Principles*, para 29.
19 *Ibid*., para 31.
20 *Ibid*.

for example, through surveillance, interception, collection and retention of data, is an additional means to counter violent extremism and terrorism online.

Article 17 of the ICCPR stipulates that while privacy is not an absolute right, it must be protected against unlawful or arbitrary interference. Specifically, unlawful interference occurs where the interference is outside the scope envisaged by the law. Arbitrary interference should also be avoided, thus even where the law provides grounds for interference, this must be reasonable, necessary and proportionate in the particular circumstances. Any storage of a person's information, whether by public authorities or private actors, should be regulated by law. The UN Human Rights Committee's General Comment No 16 recommends that states should ensure that "information concerning a person's private life does not reach the hands of persons who are not authorized by law to receive, process and use it, and is never used for purposes incompatible with the Covenant."[21]

The Council of Europe Guidelines on Human Rights and the Fight against Terrorism state, with regard to the collection and processing of personal data within the context of the fight against terrorism, that the collection and the processing of personal data by any competent authority in the field of state security may interfere with the respect for private life only if the mechanisms for the collection and processing:

(i) are governed by appropriate provisions of domestic law;
(ii) are proportionate to the aim for which the collection and the processing were foreseen;
(iii) may be subject to supervision by an external independent authority.[22]

Non-Binding International Instruments

Besides international law obligations provided by human rights treaties and customary international law, there are also non-binding instruments that can guide policy-makers and other stakeholders in the legislative process on preventing and countering violent extremism and terrorism online. The complexity of the issue has led to a number of multi-stakeholder initiatives which aim to design norms for content moderation online. Therefore, any legislative action on countering violent extremist and terrorist content online should take these new developments into account.

The *Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence*[23] annexed to the Report of the United Nations High Commissioner for Human Rights on the respective expert workshops, proposes a six-part threshold test for expression to be considered criminal. The six factors which should be considered are:

1. the context in which the statements were made;
2. the status or position of the speaker in society;

21 UN Human Rights Committee, *CCPR General Comment No. 16: Article 17 (Right to Privacy). The Right to Respect of Privacy, Family, Home and Correspondence, and Protection of Honour and Reputation*, para 10, 08 April 1988.
22 Council of Europe Guidelines on human rights and the fight against terrorism. 11 July 2002.
23 UN Human Rights Council, *Annual report of the United Nations High Commissioner for Human Rights – Report on the expert workshops on the prohibition of incitement to national, racial or religious hatred* (A/HRC/22/17/Add.4), 11 January 2013.

3. the intention of the speaker;
4. the content and form of the speech;
5. the extent to which the speech is public and disseminated;
6. the likelihood and imminence that the speech would incite a criminal act.

In 2019, a number of governments and ICT companies adopted the 'Christchurch Call to Action To Eliminate Terrorist and Violent Extremist Content Online'.[24] The Christchurch Call commits government and ICT companies to a range of measures to prevent and counter violent extremism and terrorism online. These measures include developing tools to prevent the upload of terrorist and violent extremist content; countering the roots of violent extremism; increasing transparency around the removal and detection of content; and reviewing how companies' algorithms direct users to violent extremist content.[25] At the G20 Osaka Summit in 2019, all G20 leaders adopted the G20 Osaka Leaders' Statement on Preventing Exploitation of the Internet for Terrorism and Violent Extremism Conducive to Terrorism. This statement urged online platforms to step up the ambition and pace of their efforts to prevent terrorist content from being streamed, uploaded or re-uploaded, and committed to continue working together to tackle this challenge.[26]

The Global Network Initiative, a multi-stakeholder initiative, developed a *Policy Brief on Extremist Content and the ICT Sector* which identifies a number of recommendations for governments as well as ICT companies of practices that should be avoided.[27] For example, there should be no restrictions on reporting or commentary by journalists and media outlets on terrorist groups or acts of terrorism, and law and policies should distinguish between speech aimed to incite terrorist acts and speech which debates, discusses or reports on such acts.[28]

The Camden Principles on Freedom of Expression and Quality ("The Camden Principles")[29] prepared by Article 19, elaborate on the relationship between freedom of expression and equality issues. The Camden Principles assert that the relationship between the freedom of expression and equality is mutually supportive and reinforcing, and set out recommendations on how to resolve tension between them.

B: Policy Design

In developing and adopting or updating legislation and policies to address the availability and accessibility of violent extremist and terrorist content online, the following elements discussed below should be included.

24 See https://www.christchurchcall.com.
25 Rt Hon Jacinda Ardern, *Christchurch Call to eliminate terrorist and violent extremist online content adopted*, 16 May 2019.
26 *G20 Osaka Leader's Statement on Preventing Exploitation of the Internet for Terrorism and Violent Extremism Conducive to Terrorism (VECT)*, 2019.
27 Global Network Initiative, *Extremist Content and the ICT Sector, A Global Network Initiative Policy Brief*, November 2016.
28 *Ibid.*, 4.
29 Article 19, *The Camden Principles on Freedom of Expression and Equality*, April 2009.

Human Rights Guarantees

A key issue present in most national legislation is the absence of or weak emphasis on human rights in general and the right to freedom of expression in particular. Laws allowing the blocking or removal of violent extremist and terrorist content online should be adopted and enforced in compliance with international human rights law. In practice, this may be achieved by the adoption of clear provisions on content removal based on the conditions of Article 19(3) of the ICCPR, while leaving sufficient flexibility for the provisions to remain applicable in light of technological developments.

Adopting or Updating Relevant Legislation

Existing national legislation on violent extremism and terrorism often does not reflect current technological realities, as it predates the advent of cyberspace. Preventing and countering violent extremism and terrorism online without adequate legislation may in return lead to practices that violate international human rights law and produce only limited results; on the contrary, having concrete legal provisions on the matter enables state institutions to apply the law with higher precision and thus limit the space for potential human rights violations. In addition, it has become widespread practice for law enforcement agencies to request the removal of violent extremist or terrorist content online based solely on the terms of service of a specific online platform. This may result in non-compliance by state authorities with the principle of legality, which requires every act of a state organ to be grounded in an explicit legal provision in force. Therefore, it would be desirable that states either adopt laws and regulations on addressing the availability and accessibility of violent extremist and terrorist content on the Internet, or update existing legislation with specific content-based elements.

Drafting Legal Provisions

Non-existent, vague or overly broad definitions of violent extremist and terrorist content in national laws may lead to over-zealous content removal practices based on political, religious or ideological reasons. This risk can be mitigated by the adoption of precise definitions of violent extremist and terrorist content online that is to be blocked or removed. In addition, expert training on the characteristics and distinctive features of such content should be provided to the members of law enforcement, judiciary and other relevant authorities in the field.

Independent (Judicial) Review and Appeals Processes

Various pieces of national legislation authorize law enforcement organs to refer the content that they assess to be unlawful to internet service providers (ISPs) or directly to online platforms (sometimes referred to as Content Service Providers – CSPs) leaving the final decision to the platforms or ISPs concerned. Such practice presents a potential threat to the full enjoyment of human rights by individual users to the extent no independent oversight is monitoring the procedure. This may lead to content removal practices which could seriously infringe upon human rights of persons whose content has been blocked or removed. For this reason, all state legislation on the matter should contain rules laying down the procedures to follow by state organs requesting the blocking or removal of online content as well as the rights and obligations of ICT companies as the addressees of the requests.

Furthermore, it is necessary to distinguish between the decisions on content removal by the organs of a state (notably, law enforcement and the judiciary) and those of ICT companies. For the former, a judicial decision may be indispensable as states are the primary bearers of human rights obligations and should take all necessary precautions to prevent arbitrary interference with the human rights of their citizens. With regards to blocking, which the OSCE defines as "an activity which is used to prevent access to Internet content or websites including social media platforms", the OSCE Guidebook on "Media Freedom on the Internet" recommends that policymakers "rely on blocking only within a strict legal framework with regards to content identified as illegal by the courts of law."[30]

The state's decisions also need to be subject to an appeals process if users consider their human rights have been unlawfully restricted. On the other side, ICT companies should follow a due diligence approach by establishing independent review mechanisms to allow users to challenge decisions regarding content removal based on national legislation or terms of service.

Mechanisms for Enforcement

Law enforcement actions related to preventing and countering violent extremism and terrorism online should first apply less intrusive measures such as the flagging of specific content as violent or extremist to ICT companies. Measures such as the blocking of entire websites and platforms should be kept as a last resort; as noted in the OSCE guidebook mentioned above, "blocking is not an effective method to address problems associated with Internet content and could have serious side effects including over blocking."[31]

In general, fines may be necessary to enforce legislation and ensure compliance. However, in the field of content regulation, potentially high fines, particularly when combined with loosely defined obligations, may be perceived by the ICT companies as an incentive to block or remove legal content to minimize or avoid the risk of a fine, thereby resulting in arbitrary and over-zealous restrictions of the freedom of expression. High fines can also endanger the very existence of smaller ICT companies on the market. Therefore, fines should always be proportionate and imposed in relation to clearly defined obligations.

Case Study: Germany's Network Enforcement Act

The German Network Enforcement Act (Netzwerkdurchsetzungsgesetz – "NetzDG") came into force in October 2017 (with a transitional period until 01 January 2018). It obliges social media networks to remove or block access to content that is manifestly unlawful within 24 hours of receiving the complaint, or 7 days in case of non-manifestly unlawful content (the law does not define the features of manifestly unlawful content). The network has to retain unlawful content as evidence and store it for a period of 10 weeks. Consistent or systemic noncompliance can lead to fines of up to 50 million Euros.

30 Organization for Security and Co-operation in Europe (OSCE), Media Freedom on the Internet: *An OSCE Guidebook*, 09 March 2016.
31 *Ibid.*

The law sets out a number of transparency mechanisms, namely the obligation to provide users with an easily recognizable, directly accessible and permanently available procedure for submitting complaints about unlawful content; the obligation to notify the person submitting the complaint and the user concerned about any decision, while providing the complainant as well as the author of the content with reasons for the final decision; and the obligation to produce half-yearly reports on the handling of complaints for social networks receiving more than 100 complaints per year.

Equally, oversight mechanisms include the obligation to monitor the handling of complaints via monthly checks by the management of the social network; and the oversight of the procedure by an agency tasked to do so by the German Federal Office of Justice. The law also grants the user the opportunity to respond to the complaint before the decision is rendered by the social network if the unlawfulness of the content is dependent on the falsity of a factual allegation or factual circumstances. Finally, it requires the administrative authority wishing to issue a decision (notably the decision to issue a fine) relying on the fact that content which has not been removed or blocked is unlawful to first obtain a judicial decision establishing such unlawfulness.

NetzDG has drawn a varied response. While some research has indicated it did not lead to an over-removal of content, the UN Special Rapporteur on the Freedom of Expression has raised concerns that the strict deletion periods and high fines may be disproportionate and potentially result in the removal of lawful content, and that there was a lack of judicial oversight over the removal and deletion of content by social media companies.[32] Similar concerns were expressed by eight out of ten experts invited to a hearing on the draft law.

At present, all major social media networks (Facebook, Twitter and YouTube) review the complaints filed under NetzDG first under their respective community standards. If there is a violation the content is blocked globally. If no violation is found the complaint is assessed with regards to NetzDG; in case of unlawfulness under this legislation the access to the content is blocked for Germany only.

32 Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Letter reference OL DEU 01/2017, 01 June 2017.

Content-Based Responses:

2. Development of Transparency and Accountability Mechanisms

This Chapter seeks to support policy-makers and practitioners in developing transparency and accountability mechanisms in preventing and countering violent extremism and terrorism on the Internet. These mechanisms relate to: a) providing due process to allowing an individual to challenge a content removal decision, and to b) informing the general public on the content referral and removal practices and methods by private companies and governments alike.

The Chapter aims also to elaborate on the importance of putting in place monitoring and evaluation frameworks to promote effective content referral and removal practices and prevent unintended consequences. This Chapter is divided into three sub-sections: Transparency and Accountability Mechanisms; Monitoring & Evaluation of Content-based Responses; and Automated Processes.

Relevant Good Practices from the London-Zurich Recommendations

Good Practice 4: *To develop, in collaboration with other relevant stakeholders, a common monitoring and evaluation framework that promotes transparency and facilitates greater understanding of the impact of responses.*

Good Practice 7: *To adopt laws, regulations, and policies that address the availability and accessibility of violent extremist and terrorist content on the Internet.*

Good Practice 8: *To take into account any applicable existing international standards and/or principles when addressing the availability and accessibility of violent extremist and terrorist content on the Internet and social media platforms.*

Good Practice 10: *To provide reference to the pertinent laws and regulations that motivate such referrals of relevant content to the ICT industry.*

Good Practice 11: *To acknowledge the role of the ICT industry in effectively addressing the availability and accessibility of violent extremist and terrorist content on the Internet and social media platforms.*

Good Practice 12: *To monitor and evaluate the application of automated processes that are employed to limit the re-dissemination of existing and/or already identified violent extremist and terrorist content Online.*

INTRODUCTION

While Chapter 1 set out the development and adoption of content-related legislation and policies, the present chapter focuses on transparency and accountability in their implementation.

In some countries, law- and policy-makers have substantially increased the legal, human and financial resources states have at their disposal for content analysis, referral, and removal. Additionally, social media companies have come under increasing public, economic, and political pressure to prevent violent extremist and terrorist groups from using their services and platforms, to which they have responded by collaborating with governments and civil society alike to moderate and remove violent extremist and terrorist content. Finally, private companies are increasingly required by some governments to remove unlawful content within a limited time period.[33] These different elements may lead to increased risks such as incidentally and unintentionally stifling online speech (so-called "chilling-effect" on the freedom of expression). Transparent and accountable content moderation, referral, and removal are thus vital to protect the human rights of lawful Internet users.

While removing any type of content – whether through upload filters, automated decision-making, or flagging – is a restriction on the freedom of expression, certain types of expression may be lawfully restricted by governments if the requirements elaborated in Chapter 1 are met. Nevertheless, transparency is vital for ensuring that the right to freedom of expression is respected. While governments have a primary role in strengthening transparency and accountability in content-based responses, ICT companies also have a vital role to play in strengthening transparency.

A: Transparency and Accountability Mechanisms

Content Monitoring

Monitoring of online content is an additional means to counter violent extremism and terrorism online: this can include surveillance, interception, collection and retention of data. Even if increased monitoring of online content can help detect terrorist and violent extremist sympathizers, recruiters or terrorist plots, rights enjoyed offline must also be protected online, including the right to privacy and the freedom of expression.[34] As outlined in Chapter 1, states should not adopt laws and policies that prohibit encryption and anonymity or weaken available digital security tools.

Government monitoring of online content can take a variety of forms. For example, in Sweden, the *Swedish National Security Service* (Säkerhetspolisen[35]) regularly monitors websites that might contain terrorism-related messages. The Security Sector Act (Förordning (2002:1050)

33 See Chapter 1 case study on Netzwerkdurchsetzungsgesetz; French National Assembly, *Draft Law to Fight Hate on the Internet*, N° 1785, 20 March 2019; and European Commission, *Proposal for a Regulation of the European Parliament and the Council on Preventing the Dissemination of Terrorist Content Online* (COM(2018) 640 final), 12 September 2018.

34 See UN Human Rights Council Resolutions on *The Promotion, Protection and Enjoyment of Human Rights on the Internet*: 20/8 (A/HRC/20/L.13), 05 July 2012, and 26/13 (A/HRC/26/L.2), 26 June 2014.

35 https://www.sakerhetspolisen.se/

med instruktion för Säkerhetspolisen) regulates the Security Service. However, it does not contain any provisions that specifically pertain to the monitoring of websites. If the Security Service detects content it deems illegal, it may refer the content to the ICT company on whose services the content appears and initiate preliminary investigations but is not authorized to take steps to remove any content.[36]

In Australia, the eSafety Commissioner and its Cyber Report team[37] investigate complaints about prohibited material that, for example, promotes, provides instruction in or incites in matters of crime or violence, or advocates the doing of a terrorist act. The Cyber Report team assesses reported online content against the National Classification Scheme (the Scheme) and other relevant laws, prioritizing serious material that, for example, can be considered pro-terrorist content and content that is 'abhorrent violent material'. The latter is content depicting murder, terrorism leading to death or serious injury and other violent crimes, if recorded by the perpetrator or their accomplices. These types of content are likely to be Refused Classification under the Scheme, and will be considered prohibited if hosted in Australia. Once the material is assessed, the Cyber Report team may notify it to the relevant hosting provider for takedown. A takedown notice issued by Cyber Report against an Australian hosting service is legally enforceable, and serious penalties exist for non-compliance.

In Switzerland, the Swiss Cyber Cybercrime Coordination Unit (CYCO), located at the Federal Office of Police, both actively searches the Internet for illegal content and receives corresponding reports. After assessing the respective content and securing relevant data, CYCO refers the case to the relevant law enforcement agencies.[38]

Increasing Transparency & Accountability

Transparency and accountability are vital to anticipate and mitigate potential negative implications of content-based responses in preventing and countering violent extremism and terrorism online, particularly in the context of content monitoring online and its potential impacts on the enjoyment of human rights.

Governments are encouraged to reference the relevant national legislation that forms the basis for a referral when requesting removal of content from ICT companies. Such transparency in decision-making processes strengthens trust in and between the respective stakeholders. In this respect, governments are encouraged to make pertinent laws and regulations accessible.

Strong transparency and accountability mechanisms should also provide individuals with access to remedy when their content was removed inaccurately, particularly if the rights to freedom of expression or privacy have been infringed by states and non-state actors. The UN Guiding Principles on Business and Human Rights reiterate the principle of access to remedy as a part of a state's duty to "take appropriate steps to ensure, through judicial, administrative, legislative or other appropriate means, that when such abuses occur within their territory and/

36 Swiss Institute of Comparative Law, *Comparative Study on Blocking, Filtering and Take-Down of Illegal Internet Content*, p. 671, 20 December 2015.
37 See https://www.esafety.gov.au
38 See https://www.cybersecurityintelligence.com/cybercrime-coordination-unit-switzerland-cyco-2085.html

or jurisdiction those affected have access to effective remedy."[39] Such remedy mechanisms may include obligations to provide access to "remedies and complaint mechanisms to ensure that users can challenge the removal of their content" as proposed by the European Commission.[40]

The European Commission commissioned the Institute for Human Rights and Business, and civil society organization Shift to develop an ICT-sector specific guide on the corporate responsibility to respect human rights based on the UN Guiding Principles on Business and Human Rights. The objective of the guide is to support ICT companies in translating the principles identified in the UN Guiding Principles into the ICT's own systems and cultures.[41] The guide includes concrete guidelines on how companies can develop systematic internal processes to better equip themselves to adequately and quickly handle government requests for data and/or content removal while respecting human rights. It also advocates for and provides guidance on how ICT companies can communicate these efforts effectively and transparently.

The *Santa Clara Principles on Transparency and Accountability in Content Moderation Practices,* developed by a number of academics and non-profit organizations including the Electronic Frontier Foundation, ACLU, and the Center for Democracy and Technology, advocate that companies publish the number of posts removed and accounts permanently or temporarily suspended, that companies provide notice to each user about the reason for account removal or suspension, and that companies provide meaningful opportunity for timely appeal.[42]

Case Study: Ranking Digital Rights

Ranking Digital Rights (RDR) is a non-profit research project located within New America's Open Technology Institute. RDR publishes an annual index on ICT companies' commitments and policies affecting users' freedom of expression and privacy, based on international human rights law. The RDR index hence offers clear standards to follow for companies that are committed to respecting freedom of expression and privacy as human rights and promotes transparency and accountability amongst companies through its publicly available assessments.

In 2019, the RDR Corporate Accountability Index[43] ranked 24 companies on 35 indicators that looked at "companies' governance mechanisms to identify and prevent potential threats to users' human rights, plus disclosed policies affecting users' freedom of expression and privacy."[44] Despite an improvement by companies that had been evaluated previously, there remained issues with transparency about how content removal is carried out. The Index also found that companies still failed to offer appropriate grievance and remedy mechanisms that help reporting and remedying

39 *UN Guiding Principles on Business and Human Rights*, 2011.
40 European Commission, *Proposal for a Regulation on Preventing the Dissemination of Terrorist Content Online*, 2018.
41 Shift and Institute for Human Rights and Business, *ICT Sector Guide on Implementing the UN Guiding Principles on Business and Human Rights*, European Comission, June 2013.
42 *The Santa Clara Principles On Transparency and Accountability in Content Moderation*, 02 February 2018.
43 Ranking Digital Rights, *2019 Ranking Digital Rights Corporate Accountability Index*.
44 See https://rankingdigitalrights.org/about/our-work.

harms. Nevertheless, it should be noted that member companies of the Global Network Initiative (see below) were scoring higher on the index than non-members.

RDR also includes specific recommendations for companies and for governments at the end of their report.[45] Governments may also refer to the RDR reports to better understand how companies are performing against provisions in international human rights law.

Case Study: Global Network Initiative

The Global Network Initiative (GNI) is a multi-stakeholder platform that aims to protect and advance freedom of expression and privacy in the ICT sector. The GNI Principles, which all GNI company participants commit to implementing, provide an evolving framework for responsible company decision making in support of freedom of expression and privacy rights.[46] As an increasing number of companies join the GNI, the principles are hoping to take "root as global standard for human rights in the ICT sector",[47] hence furthering human rights along with the transparency and accountability of ICT companies.

Every two years, companies that participate in GNI are independently assessed on their progress in implementing the GNI principles. The assessment aims to evaluate that companies are "making good faith efforts to implement the GNI Principles with improvement over time." As such, companies are assessed against their own prior performance. The assessments evaluate a company's systems, policies and procedures along with a small number of case study assessments on how a company handled specific incidents and how their response could be improved. The assessments are conducted by a number of independent institutions[48] that are accredited by the GNI's multi-stakeholder board according to their independence and competency criteria.[49]

Transparency Reports

While governments have a primary role in promoting transparency and accountability in content-based responses to prevent and counter violent extremism and terrorism on the Internet, such as by providing a reference to a pertinent law or criminal code, when referring certain content to ICT companies for their assessment, ICT companies can also contribute to strengthening transparency and accountability in this regard.

45 *Recommendations for governments*, in: Ranking Digital Rights, *2018 Corporate Accountability Index*.
46 Global Network Initiative, Principles on Freedom of Expression and Privacy.
47 See https://globalnetworkinitiative.org/about-gni.
48 See https://globalnetworkinitiative.org/independent-assessors.
49 Global Network Initiative, *GNI Independence and Competency Criteria*. Updated August 2018.

Issuing transparency reports that give insights into how companies have dealt with content removal on their services can be an important step for the public to better understand the extent of content removals and also what type of content has been removed. To name one example of government practice, the German Network Enforcement Act ("Netzwerkdurchsetzungsgesetz from September 1 2017 (BGBl. I S. 3352), which is further elaborated on in Chapter 1, requires social media platforms that receive more than 100 complaints per calendar year to publish a bi-annual report on how complaints were handled. It also obliges companies to notify the person submitting the complaint and the user about any decision taken, while providing the user with reasons behind the final decision. Similar provisions also exist in the proposed Regulation on Preventing the Dissemination of Terrorist Content Online by the European Commission. Such requirements provide mechanisms for promoting greater transparency with regard to how companies are conducting content regulation on social media platforms.

Case Study: Twitter

According to Twitter's rules and policies, "transparency is vital to protecting freedom of expression".[50] In this respect, Twitter publishes biannual transparency reports intended to highlight trends and provide an open exchange of information.[51]

Twitter's policy is generally to notify users as soon as possible of requests for their Twitter or Periscope account information, which includes a copy of the request, unless Twitter is prohibited from doing so. In line with Twitter's Privacy Policy, it may also disclose account information to law enforcement in response to a valid emergency disclosure request (e.g. 18 U.S.C. § 2702(b)(8) or Section 8 of Irish Data Protection 1988 and 2003).

Twitter also works together with a number of organizations, such as the Parle-moi d'islam (FR), Imams Online (UK), or True Islam (US) to counter violent extremism on its platform. According to Twitter's policy on law enforcement support, Twitter responds to valid legal processes issued in compliance with applicable laws[52] and developed its own Legal Request Submission site for law enforcement agencies.[53]

50 See https://help.twitter.com/en/rules-and-policies/tweet-withheld-by-country.
51 See latest version https://transparency.twitter.com/en/information-requests.html#information-requests-jul-dec-2018.
52 See https://help.twitter.com/en/rules-and-policies/twitter-law-enforcement-support.
53 See https://legalrequests.twitter.com/forms/landing_disclaimer.

Case Study: Open Technology Institute's Transparency Reporting Toolkit

The Open Technology Institute located within New America, a think tank based in the United States, publishes transparency reporting toolkits wherein they assess best practices for company transparency reporting and provide an overview over which indicators the most prominent ICT companies are reporting on.[54]

While acknowledging that public pressure for increased transparency has led to an increase in the frequency and depth of transparency measures by ICT companies, the Open Technology Institute advocates that more standardization in how companies report on content removal as well as further granularity of what data is reported remains of need. The lack of consistency in metrics and reporting standards poses obstacles to sector-wide and cross-company comparisons to evaluate the impact of content removal on the availability and dissemination behavior of violent extremist online. While some metrics vary between platforms because of the different types of content they host, the Open Technology Institute advocates that a uniform set of metrics that can be applied as appropriate is still helpful and necessary for cross-company comparisons and impact evaluation. Moreover, reporting on removals by companies based on violation of their own terms of service or content guidelines remains infrequent and inconsistent, although Facebook, Google and Twitter – and to a lesser extent Microsoft – have begun reporting on this indicator in 2018.

To highlight this issue, the Electronic Frontier Foundation launched their TOSsed out project in May 2019 that collects some of the content that was taken down due to the enforcement of Terms of Service rules by platforms, which it perceives to be unevenly and unfairly enforced and insufficiently transparent.[55]

B: Monitoring & Evaluation of Content-based Responses

Monitoring & Evaluation Frameworks

Demonstrating impact is crucial to ensuring the legitimacy and efficacy of actions taken to prevent and counter violent extremism and terrorism online. Continuous monitoring and evaluation of content-based responses allows for informed law- and policy-making on referral and removal processes, notably by improving content targeting and identifying and addressing human rights risks as they arise.

The fact that currently empirical research and data are lacking with regards to content-based responses and their effectiveness is particularly concerning as it means governments and ICT

54 New America Open Technology Institute, The Transparency Reporting Toolkit: Content Takedown Reporting, last updated 25 October 25, 2018.
55 See https://www.eff.org/tossedout

companies could be misallocating valuable financial and human resources and programs, which in return could bring about unintended and even detrimental consequences for human rights. Governments are thus encouraged to learn from existing monitoring and evaluation frameworks from other sectors, including public health and commercial advertising and marketing, where applicable.

Setting up Monitoring & Evaluation Frameworks

Monitoring and evaluation frameworks are key to effective and efficient of content-based responses to violent extremism and terrorism online and should be integrated into corresponding legislation and policies as well as their concrete implementation from the start.[56]

Governments are encouraged to develop, in collaboration with a range of stakeholders including the ICT industry, civil society, and academic institutions, realistic ways and means to measure the impact of legislation, policies, and programs. This means that clearly defined impact objectives for specific content-based responses (possible undergirded by Theories of Change that are outlined in respective policies) and objective baselines to assess impact should be established from the onset. Monitoring and evaluation should be based upon data, definitions, methodologies, and indicators of success that are openly communicated, consistent, and comparable.

Quantitative metrics and tools

With the vast amount of data available, it is more imperative than ever that monitoring and evaluation frameworks are focused on a determined set of data to be able to extract any meaningful information for law- and policy-makers. Furthermore, data and information collection processes should be set up in a way that protects human rights, such as the rights to freedom of expression and privacy, as well as equal protection of the law without any discrimination, which might limit the state's ability to gather certain data or at least utilize it.

Possible key elements for monitoring and evaluating content-based responses – not all of which might be available to a specific state – include:

- Number of referrals for removals and the specific grounds of referral;
- Basis of the referral (national legislation and/or terms and conditions of ICT companies);
- Nature of the natural or legal person submitting a referral, such as government institutions (possibly disaggregated between IRU and non-IRU); the judiciary; civil society; ICT companies; and individual citizens;
- Channels used for referral (IRU referral, tools available to trusted flaggers, public forms);
- Number of referrals aggregated for each respective ICT company;
- Time it took for the respective ICT company to review content referred to it;
- Number and percentage of referrals that resulted in removals;
- Where possible and appropriate, number of times content that was subsequently removed was viewed resp. engaged with and how long it was online before removal.

56 See, for instance, the European Commission's *Proposal for a Regulation on Preventing the Dissemination of Terrorist Content Online*, in particular Articles 21 and 23 on monitoring and evaluation.

- Number of removals overall and the specific grounds of removal;
- Number of removals that were appealed against and the specific appeals procedures used (judiciary/administrative/company grievance mechanism);
- Number of appeals accepted resp. rejected;
- If applicable, overview of sanctions, notably financial and penal.

Monitoring & Evaluation Tools and Capacities

The quantitative performance of content-based responses can be tracked by government/judicial general monitoring mechanisms as well as by analytics tools, notably those of ICT companies.

Referrals and removals based on judicial orders should appear in the information systems normally used by the judiciary, and government should be able to access them as is the case for other judicial proceedings. The government should define adequate ways to regularly receive updated data from ICT companies on referrals received and removals carried out under their respective terms of services, and ICT companies should comply with these government requests as long as they respect international human rights law and national legislation. IRUs will also dispose of a system to log their referrals and those should be drawn upon by the government as well.

Even with the focus discussed above, the amount of data introduced into the monitoring and evaluation system will still be substantial. Increasing government capacity to crunch significant amount of data and visualize complex data-sets can yield significant added value to constantly adapt content-based responses and the corresponding human rights safeguards. Governments can also choose to provide funding for projects at academic institutions and civil society so that those can develop innovative monitoring and evaluation approaches.

Qualitative metrics and tools

The aforementioned quantitative approaches are likely to reach a substantial array of information. Nevertheless, it is likely that combining them with qualitative elements will yield far superior evaluation results. To give one example, law enforcement specialists conduct qualitative assessments at Joint Assessment days with IRUs precisely to assess the prevalence and patterns of violent extremists' and terrorists' use of a specific platform.[57] Additional approaches could e.g. be (semi-)structured interviews with persons producing, sharing or liking violent extremism and terrorism online as well as understanding – again through personal interaction – how wider Internet users are exposed to and deal with this type of content.

Qualitative evaluation not only provides a logical frame to often overwhelming arrays of data, but also increases the transparency among participating actors on the specific practices – the qualitative review process itself becomes part of reflecting on and optimizing content-based responses.

57 Europol, *Referral Action Day with six EU Member states and Telegram*, 05 October 2018.

The Added Value of Transparent Monitoring & Evaluation

Monitoring and evaluation results should be posted publicly, whenever possible, so that non-governmental stakeholders including the ICT industry, civil society organizations, and academic institutions, can review and analyze them and make suggestions for potential revisions. A further step could be to provide even the data itself through so-called open data-sets; having those data-sets reviewed by non-governmental stakeholders can potentially provide even more in-depth insights of relevance for the strengthening of the monitoring and evaluation frameworks.

As even fully functioning institutions face the challenge of inherent analytical biases, where possible monitoring and evaluation should be carried out by autonomous government bodies (such as national statistics offices) or at least by relying on their technical expertise in setting up such monitoring and evaluation systems. Additionally, regular independent evaluations both of the systems and of samples of the collected data can help provide fresh perspectives that contribute to strengthening the monitoring and evaluations frameworks overall.

Monitoring & Evaluation Challenges

Monitoring and evaluating the impact of content-based responses comes with a set of inherent challenges. First, it is necessary, but difficult, to assess whether the removal of content leads to the migration of violent extremist and terrorist content to other platforms that may be less regulated. The different experiences between platforms in content removal further indicates the necessity in pursuing a whole-of-industry approach that evaluates the overall decrease rather than a 'success' on a single platform that may lead to an increase in terrorist content on other platforms, pushing it to smaller, less-regulated platforms or more encrypted channels.

Careful evaluation design can address some of these challenges. Nevertheless, comprehensive monitoring and evaluation frameworks should also be transparent about the limitations of the methods employed, the data that can be collected (both in technical terms and regarding data privacy requirements), and the evaluation results that can subsequently be gathered from them.

C: Automated processes

Minimizing Risks Associated With Automated Processes

As more companies develop and make use of automated processes that accelerate the identification and removal of content, the role of ICT companies in promoting effective transparency and accountability mechanisms is becoming more relevant, particularly due to the risk to the right to freedom of expression these automated processes can pose. ICT companies should ensure that automated processes are efficiently and effectively reviewed and that proper appeals mechanisms are put in place.

Automated processes can seriously infringe upon the human rights of the users concerned if the content is blocked or removed. For instance, after YouTube introduced a new technology to automatically flag and remove content violating its Terms of Service, human rights activists

complained that thousands of videos documenting alleged war crimes uploaded to YouTube were removed because the automated processes employed assessed them as violating content guidelines.[58]

Examples of Automated Processes to Detect and Remove Violent Extremist and Terrorist Content

Automated processes have been increasingly used by large social media companies to detect, flag and/or remove content on their platforms; this particularly applies to Facebook, Twitter and YouTube. Each platform uses different types of automated processes.[59]

Facebook: Facebook uses machine learning to assess Facebook posts that may signal support for Daesh/ISIL or al-Qaeda. The tool generates a score indicating the likelihood that the post violates Facebook's counter-terrorism policies that helps its team of reviewers. Additionally, Facebook has started to apply artificial intelligence (AI).[60] In particular, AI is used to route a newly uploaded piece of content to a human reviewer, to identify clusters of pages, posts, groups or profiles with terrorist content, as well as to match photos and videos against an existing database. According to Facebook's Counterterrorism Policy Manager, Facebook is in its early stages to develop text-based AI. Moreover, Facebook is using AI to reduce the time period that terrorist recidivist accounts are available on Facebook.[61]

Twitter: Twitter increasingly focuses on proactively identifying problematic accounts and behavior on its platform.[62] In a report, it is noted that 91% of a total of 205,156 accounts suspended, were proactively flagged by internal, proprietary tools. As such, Government reports constituted less than 0.1 % of all suspensions in the reported time period.

YouTube: In 2018, YouTube stressed in a blog post that "machines are allowing us to flag content for review at scale, helping us remove millions of violating videos before they are ever viewed." According to statistics provided, from October to December 2017, YouTube removed 8 million videos, 6.7 million were first flagged for review by machines and out of those 6.7 million and 76% were removed before they were ever reviewed.[63]

58 Submission by AccessNow to David Kaye, Special Rapporteur on the promotion and protection for the right to freedom of opinion and expression in response to questions for the "Study on Content Regulation in the Digital Age", January 2018.
59 Companies included here are drawn from Nikita Malik, *The Fight Against Terrorism Online: Here's The Verdict*, Forbes, 20 September 2018.
60 *A View from the CT Foxhole: An Interview with Brian Fishman, Counterterrorism Policy Manager, Facebook*. Combating Terrorism Centre, US Military Academy. September 2017, Volume 10, Issue 8.
61 Facebook, *Hard Questions: How We Counter Terrorism*, 15 June 2017.
62 Twitter, *How Twitter is fighting spam and malicious automation*, 26 June 2018.
63 See https://youtube.googleblog.com/2018/04/more-information-faster-removals-more.html.

Human Rights & Artificial Intelligence: The Role of the State

ICT companies are increasingly using artificial intelligence (AI) for content moderation, in particular automated processes. AI systems are used to police the content posted by users online for potential violations of the terms of service. While AI systems may be very efficient in identifying terrorist or violent extremist content, its algorithms may mistakenly flag legal content as illegal, which can lead to a serious infringement upon human and other rights of the users concerned if the content is blocked or removed. Therefore, legislation is necessary to regulate the use and parameters of AI-powered tools for content moderation; this includes requirements for feedback/flagging tools for individuals that consider their content to have been unlawfully removed.[64]

The UN Special Rapporteur on Freedom of Expression recommends that "states ensure that human rights are central to private sector design, deployment and implementation of AI systems."[65] The Special Rapporteur has also noted that states can meet their human rights obligations "through legal measures to restrict or influence the development and implementation of AI applications, through policies regarding the procurement of AI applications from private companies by public sector actors, through self- and co-regulatory schemes and by building the capacity of private sector companies to recognize and prioritize the rights to freedom of opinion and expression in their corporate endeavors."[66]

Finally, automated processes should always be used alongside human review in the decision to remove content on the grounds that it promotes violent extremism or terrorism, as well as decision-making in appeals processes. National legislation can set out responsibilities in this regard.

64 See, for instance, the European Commission's Proposal on *Preventing the Dissemination of Terrorist Content Online*, in particular Articles 9 and 10 on specific safeguards related to the use of automated tools.
65 UN General Assembly, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression on Artificial Intelligence technologies and implications for the information environment*, (A/73/348), para. 63, 29 August 2018.
66 *Ibid*, para. 22.

Content-Based Responses:

3. Providing content-based responses through multi-stakeholder collaboration

This chapter seeks to provide policymakers and practitioners with practices and case studies on the collaboration of governments with ICT companies and civil society. In particular, this chapter addresses collaboration between Governments, ICT companies and civil society; collaboration between ICT companies; and collaboration between ICT companies and civil society. This Chapter is divided into two sub-sections: Multi-Stakeholder Collaboration; and Further Initiatives.

Relevant Good Practices from the London-Zurich Recommendations:

Good Practice 3: *To develop a clear strategy to tackle violent extremism and terrorism online based on a whole-of-government and a whole-of-society approach, which coordinates both content-and communications-based responses, as well as offline activities, including education and engagement of civil society organizations where appropriate.*

Good Practice 6: *To adopt a multi-stakeholder approach between Governments, the ICT industry and civil society organizations in preventing and countering violent extremism and terrorism online.*

Good Practice 9: *To develop effective collaboration, where appropriate, and promote stronger engagement by the ICT industry as well as cooperation with civil society organizations when addressing violent extremist and terrorist content on the Internet and social media platforms*

Good Practice 11: *To acknowledge the role of the ICT industry in effectively addressing the availability and accessibility of violent extremist and terrorist content on the Internet and social media platforms*

Introduction

It is clear from the mission statements of the largest ICT companies that the industry, particularly social media companies, has become an essential tool for society to access, share, and discuss information. Facebook's CEO described the company's mission as "[bringing people] closer together and building a global community."[67] VKontakte defines its mission "to connect people, services, companies by creating simple and convenient communication tools."[68] Google seeks to "organize the world's information and make it universally accessible and useful."[69] Tencent strives to "improve the quality of life through internet value-added services."[70]

Due to their effective control over a significant part of the Internet's underlying infrastructure, private ICT companies play an increasing role in any effort to counter and prevent violent extremism and terrorism in the digital era. Given their tremendous role and the transnational characteristic of the digital space, effective collaboration between all stakeholders – governments, the ICT sector and civil society – is necessary for preventing and countering violent extremism and terrorism on the Internet. However, very often these private companies face a number of challenges in co- and self-regulating their platforms, especially with regards to human rights.[71]

Increasing efforts to prevent and counter violent extremism and terrorism online which combine public, public-private and private mechanisms are symptomatic of a more fundamental shift in the way business is carried out on a global scale. In light of this trend, cooperation across a variety of stakeholders – states, business and civil society – can be seen as a pragmatic response to fill some of the governance gaps found in traditional regulatory approaches. Indeed, such initiatives aim to support effective governance by ensuring that commercial actors operate within a framework of rule of law and respect for human rights. Groups composed of diverse stakeholders can together craft better approaches and solutions than would result from the work of one stakeholder group alone.

A: Multi-Stakeholder Collaboration

A Key Role for Public Institutions in Multi-Stakeholder Collaboration

States have the primary responsibility in countering violent extremism and terrorism. As laid out in Chapter 1, law- and policy-makers are thus responsible for creating adequate frameworks, in compliance with the state's obligations under international law as well as in accordance with its national law. Governments engage with ICT companies to ensure that co- and self-regulation is consistent with international human rights law and national law.

Beyond this purely legalistic approach, governments can play an important role in coordinating and engaging with the ICT sector and civil society by creating and supporting collaborative

67 Mark Zuckerberg, Building Global Community, 16 February 2017.
68 See https://vk.com/about
69 See https://www.google.com/about/
70 See https://www.tencent.com/en-us/abouttencent.html
71 Danish Institute for Human Rights, Submission to Special Rapporteur on Freedom of Expression, 28 January 2016.

platforms. These are of particular relevance to national referral units, which search for and flag terrorist and violent extremist online contact and request the removal of content via referral processes with ICT companies. Collaborative platforms can provide valuable input to governments and contribute to fostering a more inclusive decision-making process with regards to content-based responses. Open communication channels between relevant stakeholders also help to identify and plug critical gaps in effectively preventing and countering violent extremism and terrorism online, and to defuse potential conflicts of interest. Institutionalized and coordinated efforts can also promote complementary actions of, and the channeling of human and financial resources between, the various stakeholders.

Case Study: Internet Referral Units at the EU and the National Level

The European Union's (EU) Internet Referral Unit (IRU) forms part of Europol's European Counter Terrorism Centre and comprises of a team of experts in the fields of religiously inspired terrorism, languages, information and communication technology developers and law enforcement agencies specialized in counter terrorism.[72] It started its work in 2015 and has the following mandate:

- To support the competent EU authorities by providing strategic and operational analysis;
- To flag terrorist and violent extremist online content and share it with relevant partners;
- To detect and request removal of internet content used by smuggling networks to attract migrants and refugees;
- To swiftly carry out and support the referral process, in close cooperation with the industry.[73]

The IRU is responsible for assessing online content and referring it to the respective ICT company hosting the content for removal. According to the EU IRU's transparency report in 2017, "cooperation with the private sector is fundamental in prevention".[74] Since its establishment in July 2015 until December 2017, the EU IRU has assessed 46'392 pieces of terrorist content that triggered 44'807 decisions for referral with a 92 percent rate of content removal. [75] The EU Directive on combatting terrorism provides safeguards with respect to content removal outlined in Article 21 (3): "*Measures of removal and blocking must be set following transparent procedures and provide adequate safeguards, in particular to ensure that those measures are limited to what is necessary and proportionate and that users are informed of the reason*

72 See https://www.europol.europa.eu/about-europol/eu-internet-referal-unit-eu-iru
73 *Ibid.*
74 EU Internet Referral Unit, *Transparency Report 2017.*
75 *Ibid.*

for those measures. Safeguards relating to removal or blocking shall also include the possibility of judicial redress."[76] In case the assessed content infringes Europol's mandate, the relevant content is referred to the ICT company on whose platform the content has been detected. Nevertheless, eventually it is left to the discretion of the company to remove this flagged content or not, after assessing it against its respective terms of service. The EU IRU has no legal power to request companies to take down content.

Similar referral units exist in the UK, France and the Netherlands, with statements by Europol indicating that parallel mechanisms have been established in Belgium, Germany, and Italy.[77]

To promote a coordinated approach between governments and ICT companies, the EU IRU organizes so-called joint Referral Action Days, bringing together specialized law enforcement units from multiple national IRUs and the EU IRU and ICT Companies.[78]

The increasing use of referrals by IRUs has been criticized by civil society organizations such as the Global Network Initiative (GNI) to the extent that such referrals are not accompanied with adequate access to remedy, accountability, or transparency for users and the public.[79] GNI also issued a statement about its concern that some IRUs may allow content to be flagged which may violate ICT company's terms and conditions without determining if that content violates national laws.[80]

Multi-Stakeholder Collaboration: Playing to the Strengths of Respective Stakeholders

A multi-stakeholder approach is more likely to be effective and sustainable if the stakeholders involved have a common understanding of their respective roles and responsibilities and acknowledge their own strengths and limitations. Such an approach can bring together the political, legal, societal and technical know-how and expertise necessary to effectively address the availability and accessibility of violent extremist and terrorist content on the Internet.

76 *Ibid.*
77 Europol, *Referral Action Day*, 2018.
78 Europol, *EU Law Enforcement and Google Take on Terrorist Propaganda in Latest Europol Referral Action Days*, 16 July 2018. Europol, Referral Action Day, 2018.
79 See Global Network Initiative, *Extremist Content and the ICT Sector*, 2016, and Jason Pielemeier and Chris Sheehy, *Understanding the Human Rights Risks Associated with Internet Referral Units*, Global Network Initiative, 25 February 2019.
80 Global Network Initiative, Understanding the Human Rights Risks Associated with Internet Referral Units. 25 February 2019.

Case Study: YouTube's Trusted Flaggers and YouTube Contributors Program

In 2012, YouTube developed a Trusted Flagger program,[81] which allows invited individual volunteers, government agencies and non-governmental organizations that are particularly active in flagging content violating YouTube's Community Guidelines, to access tools that help them flag content more effectively (the Trusted Flaggers program does not include the flagging of content that would require the immediate closure of the account under national law). Once content is flagged, YouTube's trained content moderation teams review the content to determine whether to remove the flagged videos or not. Since Trusted Flaggers are expected to flag with a high degree of accuracy, content they report on is reviewed with priority. They also receive access to a tool that allows for reporting multiple videos at the same time, increased visibility into the decisions taken by YouTube's team on content removal and – in the case of civil society organizations – online trainings.

As shown in YouTube's transparency report, in the period between January and March 2019, individual trusted flaggers flagged the largest amount of videos detected by humans that were subsequently removed (1,396,945 videos that were then removed, compared to 4022 by NGOs and 16 by government agencies), which represents about a sixth of the content that was taken down following an automated flag (amounting to 6,372,936 videos).[82]

B: Further Initiatives

ICT Industry- & Civil Society-Led Initiatives

Social media platforms have become essential tools for society to discuss, share, and access information. ICT companies are very often faced with co- and self-regulation challenges in relation to their platforms, particularly regarding protecting human rights such as the freedom of speech and the right to privacy. In response, ICT industry-led initiatives such as knowledge and technology sharing between companies; the creation of platforms for interactive content-moderation tools and resources; and training sessions run by larger companies for smaller ones on content removal approaches, can be effective mechanisms for preventing and countering terrorist and violent extremist content on the Internet.

ICT companies can consider introducing and implementing, on a voluntary basis, individual practices dealing with violent extremism and terrorism online, such as codes of conduct or ethics on the circulation of images, videos and other related visual information, which should also be reflect in their terms of service. This could to raise their own awareness and responsibility, and complement national legislation.

81 See https://support.google.com/youtube/answer/7554338?hl=en&ref_topic=2803138
82 See https://transparencyreport.google.com/youtube-policy/removals?hl=en

Case Study: Global Internet Forum to Counter Terrorism (GIFCT)

In 2017, YouTube, Facebook, Microsoft and Twitter founded the *Global Internet Forum to Counter Terrorism* (GIFCT), whose mission is to "substantially disrupt terrorists' ability to promote terrorism, disseminate violent extremist propaganda, and exploit or glorify real-world acts of violence using our platforms."[83] In 2019, Dropbox joined the GIFCT.

The GIFCT aims to share technology and tools, and to run training sessions for smaller companies regarding tackling terrorist content, which it does in partnership with UN-affiliated Tech Against Terrorism. For example, the GIFCT created a database, allowing companies to create "digital fingerprints" of any terrorist content posted (so-called "hash-sharing database"). In June 2019, this database contained over 200,000 hashes.[84]

Following the adoption of the 'Christchurch Call to Action To Eliminate Terrorist and Violent Extremist Content Online' in May 2019, the GIFCT has pledged to also focus on crisis response by "introducing joint content incident protocols for responding to emerging or active events like the horrific terrorist attack in Christchurch, so that relevant information can be quickly and efficiently shared, processed and acted upon by all member companies".[85]

New America has raised concerns about the GIFCT's knowledge-sharing initiatives; these do not have the capacities to clearly evaluate and monitor their success, which can crowd out the development of innovative practices by smaller companies that could work more effectively. Since GIFCT members essentially establish good practices that are shared with smaller platforms without careful and strategic evaluation, the agency of smaller platforms to develop and implement new and innovative strategies becomes limited.[86]

83 See https://www.gifct.org/about/
84 *Ibid.*
85 See Facebook, *Global Internet Forum To Counter Terrorism: An Update on Our Progress Two Years on*. 24 July 2019, and Microsoft, *The Christchurch Call and Steps to Tackle Terrorist and Violent Extremist Content*, 15 May 2019.
86 Spandana Singh, *Taking Down Terrorism: Strategies for Evaluating the Moderation and Removal of Extremist Contents and Accounts*. New America.

Case Study: Tech Against Terrorism

Tech Against Terrorism is a UN-mandated initiative and public-private partnership.[87] Tech Against Terrorism has monitored terrorists' use of the internet since 2016 and its research shows that amongst the top 50 most used platforms by terrorist and violent extremist groups, about half are small- or micro-platforms. Tech Against Terrorism aims to particularly reach out to these smaller companies as they often do not have the financial, human and technical resources to effectively prevent and counter the misuse of their platforms by violent extremism and terrorism.

Companies that join Tech Against Terrorism agree to the Tech Against Terrorism pledge,[88] which contains six simple and accessible guiding principles of best practices: respecting the freedom of expression; respecting the right of users to express diverse views and opinions; safeguarding users' privacy; providing transparency surrounding content removal; and what content is permissible, along with providing access to an appeal mechanism and committing to further collaboration. The pledge aims to act as a "starting point from which companies can build their own appropriate systems and policies." It is based on international law instruments as well as the Global Network Initiative's principles.

To support small companies, Tech Against Terrorism launched its 'Knowledge Sharing Platform' in 2017 where smaller ICT companies can access specific tools and toolkits such as sample Terms of Service and model guidelines for transparency reports to help equip them with the tools necessary to better prevent and counter terrorist and violent extremist exploitations of their services.

Case Study: INHOPE – Lessons To Be Learned For Regulating Online Violent Extremist and Terrorist Content

The International Association Of Internet Hotlines has a global presence in 43 countries and seeks to contribute to an internet that is "free of child sexual abuse and exploitation".[89] Its mission is to "strengthen the international efforts to combat child sexual abuse material".[90] INHOPE partners with a variety of stakeholders including Interpol, Europol, Twitter, Crisp Thinking, Microsoft, Google, Facebook and Trend MICRO.

87 See https://www.techagainstterrorism.org/
88 Tech Against Terrorism, *The Pledge for Smaller Tech Companies*.
89 See http://88.208.218.79/gns/home.aspx
90 See http://88.208.218.79/gns/who-we-are/our-mission.aspx

INHOPE consists of 48 hotlines that provide a mechanism to the public to report content or activity online that is suspected to be illegal. INHOPE's primary focus is child sexual abuse material but it also includes hate speech and xenophobic content online. While INHOPE does provide a definition for hate speech, it also acknowledges that hate speech is an "extremely complex" matter that is often not illegal under criminal law. Therefore, each report to a hotline concerning hate speech will be assessed against national legislation, i.e. where the respective content is hosted.[91]

Another lesson learned from INHOPE's operation is the importance of staff well-being for content moderators and the acknowledgment of the psychological toll content review of violent and terrorist content can have on reviewers. A white paper, developed and published by the French hotline Point de Contact, intends to develop a common set of best practices for the operational handling and processing of harmful and potentially illegal content that may endanger physical safety and psychological well-being of professional content reviewers.[92]

91 See http://88.208.218.79/gns/internet-concerns/overview-of-the-problem/hate-speech.aspx
92 Point de Contact of the Guide d'Usage pour la Lutte contre la Pédopornographie, *Child sexual abuse material and online terrorist propaganda Tackling illegal content and ensuring staff welfare*, 2014.

Communications-Based Responses:

4. Development, Adoption and Evaluation of Policies

This Chapter seeks to provide policymakers with good practices and case studies on the development, adoption and evaluation of impactful policies and programs regarding communications-based responses in relevant Government strategies and National Action Plans. The chapter is divided into three sub-sections: Policy Design; Monitoring & Evaluation; and Ethics & Security Risks.

Relevant Good Practices from the London-Zurich Recommendations:

Good Practice 1: *To adopt and implement law and policy frameworks at the national level to prevent and counter violent extremism and terrorism online.*

Good Practice 2: *To maintain a comprehensive understanding of the current and likely future online threats presented by violent extremism and terrorism in each national and local context.*

Good Practice 3: *To develop a clear strategy to tackle violent extremism and terrorism online based on a whole-of-government and a whole-of-society approach, which coordinates both content- and communications-based responses, as well as offline activities, including education and engagement of civil society organizations where appropriate.*

Good Practice 4: *To develop, in collaboration with other relevant stakeholders, a common monitoring and evaluation framework that promotes transparency and facilitates greater understanding of the impact of responses.*

Good Practice 5: *To strengthen international cooperation as a key component to effectively preventing and countering violent extremism and terrorism online.*

Good Practice 16: *To ensure that all campaigns are centered on an overall goal, which may be as simple as promoting dialogue and engagement; a realistic set of measurable objectives; and a robust evaluation methodology to determine impact on target audiences.*

Good Practice 17: *To be aware of and take steps to mitigate against the possible risks involved in the strategy and delivery of communications campaigns.*

INTRODUCTION

Proactive communications-based policies should be balanced with content-based responses, situated within a comprehensive online and offline approach to preventing and countering violent extremism and terrorism, and address the fundamental internal and external drivers of violent extremism conducive to terrorism.[93] In line with UNSC Resolution 2354 on countering terrorist narratives, all policies designed and adopted to counter terrorism and violent extremism must comply with States' obligations under international law, including international human rights law, and respect the rule of law, while respecting privacy and freedoms of expression, association, peaceful assembly, and religion or belief.[94]

Communications policy strategies benefit from the development and adoption of clear legal or official definitions of key terms such as "(countering) violent extremism" and "terrorism" in any national legislation, strategies or Action Plans.[95] Definitions can play an important role in shaping States' understanding of the problem, effectively delimit and target their responses, and help ensure that all stakeholders approach the challenge in a coordinated fashion. It is essential that Governments communicate the intent and content of their policies effectively in the domain of preventing and countering violent extremism and terrorism online. Online communications should complement and supplement offline messaging and activities in these areas, or there is a danger that the credibility of both Governments and their policies will be undermined.

A: Policy Design

Addressing All Forms of Violent Extremism and Terrorism

Policies for communications-based approaches should address violent extremism and terrorism in all its forms, and should emphasize that violent extremism and terrorism are not exclusive to any ethnicity, religion, nationality, or belief. Violent extremist and terrorist groups use a range of different tactics and types of content for a range of audiences, including the general public, at-risk or vulnerable audiences, and committed supporters. Violent extremist and terrorist groups are also increasingly designing tailored radicalization and recruitment content and engagement strategies to target women and girls. Holistic communications-based responses should therefore take this into account, with strategies and policies designed to prevent and counter such tactics.

93 *UN Global Counter-Terrorism Strategy Review* (A/RES/70/291), para. 39, 19 July 2016.
94 UN Security Council Resolution 2354 (2017).
95 Without precluding other definitions or terms found elsewhere, including in national law, a reference point which may be considered for what could be commonly understood as "terrorist acts" is provided by UN Security Council Resolution 1566 (2004), para. 3: "[...] criminal acts, including against civilians, committed with the intent to cause death or serious bodily injury, or taking of hostages, with the purpose to provoke a state of terror in the general public or in a group of persons or particular persons, intimidate a population or compel a government or an international organization to do or to abstain from doing any act, which constitute offences within the scope of and as defined in the international conventions and protocols relating to terrorism, are under no circumstances justifiable by considerations of a political, philosophical, ideological, racial, ethnic, religious or other similar nature [...]."

To comprehensively address the range of violent extremist and terrorist content available online, a wide variety of communications-based responses are required. For the purposes of this toolkit, communications-based responses are broadly divided into 'upstream' and 'downstream' approaches (*see also Chapter 5: Figure 1*):

- **Upstream approaches** are preventative, and aimed at broader audiences. They are intended to build resilience to violent extremist or terrorist narratives, raise public awareness of Government policies or support services, or refute disinformation through education or positive or alternative narrative responses.
- In contrast, **downstream approaches** aim to more directly rebut, refute or counter the narratives of violent extremist or terrorist groups or attempts at the justification, incitement or glorification ("*apologie*") of terrorist acts. These approaches are intended for very specific audiences, including those that are already radicalized to violence or sympathetic to online violent extremist or terrorist narratives, or that are considered particularly vulnerable or at-risk to radicalization or recruitment. Downstream approaches include counter-narrative campaigns targeted at more specific at-risk audiences, and online individualized interventions for those participating in violent extremist and terrorist communities online.

Whole-of-Society Approaches

Communications-based policies should aim to limit the impact of violent extremist and terrorist communications, as well as work to address the underlying internal and external drivers of violent extremism and terrorism. As such, preventing and countering violent extremism and terrorism through communications-based responses should not be considered purely a security issue, but instead a multi-faceted challenge requiring multi-disciplinary, multi-institutional whole-of-society approaches. Governments should play a leading role in advocating for a whole-of-society approach. Thereby, their policies and strategies should encourage relevant stakeholders, including ICT companies and civil society organizations, where appropriate, to coordinate and cooperate on communications-based approaches.

These should be designed and adopted in line with broader national strategies and policy frameworks to prevent and counter violent extremism and terrorism to ensure that online and offline efforts are harmonized. Offline efforts may include initiatives to build critical thinking, digital literacy and resilience through public awareness and education, grassroots community engagement, and other approaches that address the internal and external drivers that may lead individuals to support violent extremism and terrorism.

Researchers, academics and practitioners can also provide insights into violent extremist and terrorist communications and inform potential responses. Successful approaches will draw upon expertise across a wide variety of inter-related sectors and fields, including, but not limited to: technology, marketing, advertising, content production, communications studies, psychology, sociology, political science, education, and public policy. As well as professional expertise, the views and values of key target audiences should also be considered, and where possible specific audiences (e.g. youth, women) should be involved in the design and delivery of responses.

States should consider whether the efficiency and effectiveness of holistic approaches to prevent and counter violent extremism and terrorism online may be increased by establishing a national interagency coordinating body to orchestrate and integrate both online and offline whole-of-government initiatives and programs, refine strategies and policies, and share research and monitoring and evaluation findings.

Case Study: Canada Centre for Community Engagement and Prevention of Violence[96]

The Canada Centre was founded in 2017 and is responsible for the Government of Canada's countering radicalization to violence initiatives. The Centre's duties include developing policy guidance, promoting multi-stakeholder coordination and collaboration, targeted programming and funding, planning and coordinating research. The Centre has a particular focus on promoting community efforts, including the creation of a National Expert Committee to provide guidance and inform policies and activities. The Canada Centre's Community Resilience Fund supports prevention efforts, and to date has funded twenty-four projects worth a total of over $16 million CAD.[97]

In 2018, the Canada Centre produced *The National Strategy on Countering Radicalization to Violence*, which outlines the government's three priorities for preventing and countering radicalization:

1. Building, sharing and using knowledge;
2. Addressing radicalization to violence online;
3. Supporting interventions.

The strategy provides clear and detailed definitions for radicalization, radicalization to violence and violent extremism, recognizing that there are many factors that contribute to the process, including exposure to terrorist or violent extremist narratives on and offline. *The National Strategy* clearly states the Government of Canada's commitment to "the protection of human rights and fundamental freedoms, including Charter-protected freedom of expression and privacy rights" as well as to "diversity and political inclusion for all Canadians".[98]

The National Strategy incorporates efforts to prevent and counter radicalization are divided into three streams to address all stages of the radicalization process, from early prevention, to at-risk prevention and disengagement. Online radicalization is given particular prominence as one of three key priorities, and emphasizes the need to foster communication between the Government, civil society, technology

96 See https://www.publicsafety.gc.ca/cnt/bt/cc/index-en.aspx.
97 See https://www.canada.ca/en/public-safety-canada/news/2018/12/launch-of-national-strategy-on-countering-radicalization-to-violence-and-update-on-terrorist-threat-to-canada-terrorism-threat-level-unchanged.html.
98 See https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/ntnl-strtg-cntrng-rdclztn-vlnc/index-en.aspx.

companies and international actors, and supporting research to build the evidence base on how violent extremist and terrorist groups operate online.

The Community Resilience Fund is intended to support civil society initiatives that promote digital literacy and alternative narratives, and has funded several programs including;

- *Canada Redirect* (Moonshot CVE) to target positive, alternative content to vulnerable individuals actively searching for violent extremist material online using online advertising and video content.
- *Pushing Back Against Hate in Online Communities* (Media Smarts) to research levels of understanding of online hate speech and radicalization among secondary school students to inform schools and parents responses.
- *SOMEONE (Social Media Education Every Day) Multimedia Portal* (Concordia University) to build resilience against hate speech and radicalization leading to violence among young people by developing a series of evidence-based resources for educators, the media, Government and the general public to improve responses to these challenges in a variety of educational settings, from primary to post-secondary.[99]

Case Study: The National Counter Extremism Policy Guidelines – National Counter Terrorism Authority (NACTA) Pakistan

Whole-of-Government and Whole-of-Society Approach

The Government of Pakistan's National Counter Extremism Policy Guidelines (NCEPG) were composed following 34 rounds of meetings with 305 stakeholders, and are based 'on a whole-of-government and society approach'.[100] Stakeholders consulted included members of provincial government, academics, media representatives, religious scholars and civil society organizations. All 34 rounds of discussion were guided by the Constitution of the Islamic Republic of Pakistan to ensure a commitment to human rights, minority and marginalized communities, and women.

The policy strategy notes the integral role of survivors and formers of violent extremism in countering violent extremist and terrorist narratives and aims to support the creation of a platform for their stories. The NCEPG acknowledges the issue of finding 'credible and convincing' messengers for these narratives, focusing on messengers that have a similar background to the intended audience.

99 See https://www.publicsafety.gc.ca/cnt/bt/cc/fpd-en.aspx.
100 National Counter Terrorism Authority – Pakistan, National Counter Extremism Policy Guidelines January 2018.

Communications-Based Responses as Part of a National Strategy

The NCEPG recognize the importance of online and offline media engagement not only as a means of disseminating information, but as an active tool in communications-based counter-extremism work. This includes uses of media to help humanize the stories of victims of violent extremism and to help deconstruct violent extremist narratives.

The NCEPG document also includes recommendations for the creation of a media cell for Countering Violent Extremism by the Ministry of Information and Broadcasting in partnership with the National Counter Terrorism Authority to ensure the 'synchronization of implementation of communication strategy for preventing violent extremism in society'. This cell is intended to work in conjunction with provincial Information Departments to ensure local populations are kept up to date regarding CVE programs active in their region.

Examples of Communications Campaigns Supported by NCEPG

PurAzm Pakistan is one example of a program supported by the national policy strategy set out in NCEPG. PurAzm is a media campaign showcasing the stories of everyday Pakistani citizens as well as police officers, polio workers, hospital doctors and public officials, which aims to disseminate the narrative that Pakistanis 'reject the evils of violent extremism, and remain resilient and hopeful, despite its adverse effects'.[101]

The initiative has produced 30 short films since 2014. The PurAzm program includes the PurAzm awards, which looks to support the sustainability of the program through helping university students and young professionals to create 'original, indigenous audio-visual and text based content on the themes of Purazm Pakistan'.

The Role of Media

Comprehensive communications-based strategies and policies may also consider the potential role and impact of the media to "enhance dialogue and broaden understanding", and in "promoting tolerance and coexistence, and in fostering an environment which is not conducive to incitement to terrorism, as well as in countering terrorist narratives".[102] Governments should not enact policies that infringe on the freedom, pluralism or equality of perspectives within the media. Approaches in this area should not seek to regulate the media, with any attempts to collaborate with the media occurring on a voluntary or independent basis. Governments may also play a role in supporting the diversity of sources and promoting access to media.[103]

101 See http://purazm.gov.pk/about/.
102 UN Security Council Resolution 2354 (2017), preamble para. 13.
103 Article 19, 'Hate Speech' Explained A Toolkit, 2015.

The Council of Europe, in its *Declaration on freedom of expression and information in the media in the context of the fight against terrorism*, encourages journalists and the media to consider their role in not inadvertently bolstering the aims of terrorists. This includes not inadvertently contributing to the climate of fear that terrorism can foster, and not providing a platform to terrorists through disproportionate coverage. The Council of Europe encourages the media to consider adopting and enacting relevant best practices, where not already in place, or to adapt existing approaches to ensure the potential ethical issues raised by media reporting on violent extremism and terrorism.[104] One example of such an approach is the Code of Conduct ('*Rules of Conduct for the Mass Media in Case of a Terrorist Attack and an Anti-Terrorism Operation*') voluntarily adopted by the mass media in Russia in 2003.[105] The Code primarily focuses on best practices for media during ongoing terrorist incidents to avoid compromising the operational security or further endangering lives, but also stresses the importance of rights to free expression, and enabling public discussion on issues such as terrorism.An example of a government-sponsored program considering media's possible role in countering violent extremist and terrorist narratives is one of the programs in the U.S. Department of State's ongoing International Visitors Leadership Program (IVLP) that focused on: 'Countering Violent Extremism – Media Messaging and Strategies'. This specific project worked with journalists, experts and government officials across the globe to highlight the positive roles and responsibilities of the media (both online and in print) in supporting democracy and preventing and countering violent extremism and terrorism,[106] The project also examined the role of governments in adhering to the rule of law and enabling a free press.

International cooperation is indispensable to preventing and countering violent extremism and terrorism online given the transnational nature of both the threat and the online world. International cooperation facilitates capacity building through the sharing of good practices that contribute to ensuring that national responses to limit the impact of and counter violent extremist and terrorist propaganda – both online and offline – are complementary and sustainable.

International forums can help create synergies within the international community to maximize collective efforts, and pool expertise on preventing and countering violent extremism and terrorism online. Additionally, such forums can create an environment of mutual trust, contribute to the building of platforms for enhanced communication, and ensure the efficient and effective application of resources. Governments are therefore encouraged to continually share best practices and information about national evaluation programs and policies, and work towards shared monitoring and evaluation frameworks and metrics for success (*see B: Monitoring and Evaluation Communications-Based Responses*).

104 Cf. *Declaration on freedom of expression and information in the media in the context of the fight against terrorism*, adopted by the Committee of Ministers, 02 March 2005.
105 Anti-Terrorism Convention (Rules Of Conduct For The Mass Media In Case Of A Terrorist Attack And An Anti-Terrorism Operation, 11 April 2003.
106 Chiemelie Ezeobi, *Nigeria: Countering Violent Extremism*, allAfrica, 13 June 2018.

Case Study: The Organization for Security and Cooperation in Europe (OSCE) National Table Top Exercise on Countering the Use of the Internet for Terrorist Purposes[107]

In January 2019, the OSCE's Project Coordinator in Uzbekistan and Action against Terrorism Unit held a three day, Table Top Exercise (TTX) on Countering the Use of the Internet for Terrorist Purposes based on the GCTF's *Zurich London Recommendations*. The event was designed to continue previous work by the OSCE, expanding this to include relevant civil society actors, while also focusing on the inclusion of human rights and gender issues. The TTX aimed to employ a "whole-of-society" approach by including 45 representatives from Government, law enforcement, media, academia, youth organizations, and the ICT industry.

Organizers used a fictional case study based on real-world security trends to open the exercise and facilitate discussion and dialogue between participants, who also heard presentations from various international experts and OSCE advisers. Each day of the event was given a theme (intervention, prevention, policy development) and a Facilitator's Guide was developed to direct discussions and ensure tangible outputs.

Clear Objectives:

The event was intended to produce an *Actionable Policy Reponses and Recommendations (APRR)* paper and National Action Plan to improve the efficiency of efforts to address the threats posed by the use of the internet for terrorist purposes. The APRR summarizes the topics and issues discussed during the event into coherent themes and clear and 'actionable policy responses'. The underlying aims were to illustrate potential avenues for cooperation and collaboration between stakeholders and to ensure policy recommendations are in keeping with international laws and commitments, especially concerning human rights.

Whole-of-Society Approach:

The TTX was designed as a whole-of-society approach, including members of the Uzbek government, civil society, and the ICT industry as well as international experts. The success of the event depended upon clear communication between all parties and the daily proceedings of the table top exercise were structured to ensure communication and mutual understanding. Facilitators were directed to ask participants 'probing questions' and were provided with a set of sample questions and key issues. This framework allowed not only for successful dialogue, but for appropriate responses to the issues raised to be accommodated. For example, the event was not originally intended to focus on communications-based responses, but when it became clear that there was a significant 'lack of understanding' regarding the effectiveness and implementation of such approaches, the organizers addressed this and included the topic in the event's proceedings.

107 See https://polis.osce.org/national-tabletop-exercise-countering-use-internet-terrorist-purposes.

Sustained and effective communication is also a relevant component of each of the three themes included in the APRR, with sections two and three explicitly referencing 'collaboration' and 'strategic communication' respectively. In relation to legal issues, the findings of the discussion state that national and international laws pertaining to violent extremists' and terrorists' use of the internet 'must be sufficiently detailed' to inform citizens and help safeguard against 'arbitrary or unlawful interference with the right to privacy'. Similarly, by involving members of the media, the organizers ensured that the APRR included references for the provision of training for journalists to give 'effective media coverage of terrorist threats and attacks'.

Actionable Policy:

The completion of the APRR and the National Action Plan ensures continued collaboration between civil society, government, and ICT industry by outlining policy goals and future projects that involve representatives from each of these groups. The APRR recognizes three key themes: legal frameworks on crimes related to violent extremist or terrorist uses of the internet; public-private partnerships and collaboration with the global ICT industry; strategic communications, media, education and research. The recommendations for each theme include time-frames, an outline of the actors responsible for implementation, and a list of measureable indicators to evaluate their effectiveness.

Transparency & Acknowledgement of Risks:

Recognizing human rights issues as a guiding principle for the TTX and subsequent policy recommendations ensured that the risks of countering violent extremists' and terrorists' use of the Internet were discussed and accounted for. Event facilitators explained how poorly orchestrated campaigns can, among other things, increase the risk of radicalization or inadvertently promote singular and exclusionary interpretations of religion.

Given the technical complexity of the Internet, and its constant and rapid evolution, national strategies and policies should be designed with a clear understanding of both the opportunities provided by online communications-based responses, as well as the vulnerabilities that can be exploited by violent extremists and terrorists. Governments should therefore ensure that policies and strategies are flexibly designed based on up-to-date research, and evaluated, reviewed and iteratively updated on an ongoing basis to keep pace with changes in the online environment and the digital tactics of violent extremists and terrorists.

National Action Plans or strategies may be updated on an annual basis, yet online trends often change much more quickly. The analysis of trends in relevant online audiences, platforms and popular content, as well as understandings of Internet infrastructure and architecture (e.g. online "echo chambers" and algorithmic "filter bubbles"), on which communications-based

policies are designed, should also be regularly reviewed and iteratively updated to ensure those policies remain effective.[108]

Violent extremist and terrorist online communications that attempt to radicalize and recruit individuals to violence, and polarize communities, are typically highly adaptive and quick to exploit changes to online and cultural environments. Governments should therefore also invest in research and analytical tools to maintain comprehensive understandings of the evolving online and offline intentions and impacts of terrorist and violent extremist communications, and broader online trends in terms of relevant audiences, platforms and influencers.

B: Monitoring & Evaluation

Measurement of impact should sit at the heart of all communications approaches adopted to challenge violent extremist or terrorist communications online. Mechanisms through which Governments can measure the effects of their communications, whether positive or negative, are critical in any design for communications approaches as well as for specific campaigns. Such an approach to monitoring and evaluation will improve understandings of the long-term impact of communication-based responses online, and allow future responses, both nationally and internationally, to be adapted accordingly.

Sustained, long-term investments in monitoring and evaluation, including where appropriate through collaboration with ICT industry, academia and civil society, enable resources to be effectively allocated to more effective programs. Comprehensive, embedded, and ongoing measurement of impact will also contribute to increased transparency and accountability, by helping to identify both intended and unintended outcomes of responses.

Monitoring & Evaluation Frameworks

Due to the complexity and range of potential impacts of communications-based responses, governments should develop a common and comprehensive monitoring and evaluation framework that provides clear indicators and metrics across its variety of approaches. Demonstrating impact is crucial to ensure the legitimacy and efficacy of actions taken to prevent and counter violent extremism and terrorism online.

Monitoring and evaluation frameworks should also be designed to monitor and capture the impact of responses on a range of audiences to ensure they are non-discriminatory, and equally achieve the intended outcomes across the target audience. Given that communications-based responses to violent extremism and terrorism online remain a developing field, Governments are encouraged to learn from existing monitoring and evaluation frameworks from other sectors, including public health and commercial advertising and marketing, where applicable.

108 Online "echo chambers" describe the phenomenon where individuals are exposed to conforming ideas and opinions at the expense of alternative or dissenting views. "Filter bubbles" are likely to occur where search engines or social networks personalise search results or newsfeed content through machine-learning models and algorithms that recommend content based on an individuals' location, demographic information or past online behaviour, and are therefore more likely to agree with.

Theories of Change, Goals & Objectives

Government policies should be based on a well-defined theory of change that explains how and why any communications-based responses employed contribute to the goals and objectives of the overarching National Action Plan or strategy. The theory of change should be integral to the design and implementation of any communication-based responses, and provide a framework with which to evaluate their impact.

Starting with the desired behavioral or attitudinal impacts on the intended target audience, the theory of change should outline the steps that will be required for communications-based responses to achieve the desired outcomes and impacts, and how these will be measured. An effective and realistic theory of change relies on clearly defined key concepts. Any divergence in terms of understandings of definitions will need to be addressed in order to achieve the necessary buy-in from key stakeholders, and accurately measure impact.

An overall long-term goal and a related series of immediate objectives should be set before both the design and the dissemination stages of a campaign. This provides a series of benchmarks against which to measure impact on the intended target audience. Objectives should be clearly defined, quantifiable measures of a desired effect. They should be measurable, allowing their success to be discerned from available metrics and indicators, and realistic with respect to the resources available, as well as the performance of previous efforts.

So-called "calls to action", where a campaign asks the target audience to take a specific action in response, can be an effective method for mobilizing support and encouraging and reinforcing attitudinal or behavioral change, as well as providing a tangible metric to aid the measurement of impact. Calls to action can also be an effective means of mobilizing both online and offline support, and prevent communications campaigns from being seen as superficial or lacking depth by target audiences. Such approaches must be sustained to avoid an initial enthusiasm that later fades, leaving participants skeptical of the value of the campaign, and reducing the possibility of future mobilization.

Case Study: The Global Engagement Center, United States Department of State

The Global Engagement Center (GEC) leads the U.S. government's efforts to counter communications from international terrorist organizations and foreign states. The GEC was established by the Secretary of State in 2016, with a mission to "lead, synchronize, and coordinate efforts of the Federal Government to recognize, understand, expose, and counter foreign state and non-state propaganda and disinformation efforts aimed at undermining United States national security interests."[109] The GEC expanded upon the earlier U.S. interagency initiative, the Center for Strategic Counterterrorism Communications (CSCC), which also was housed within the Department of State.

109 See https://www.state.gov/about-us-global-engagement-center/.

Clear Strategy:

An interagency approach enables the GEC to coordinate communications efforts without duplication across the U.S. Government. Coordination with national security departments helps to inform the objectives of the GEC's activities with up-to-date insights and intelligence. This interagency communication ensures that GEC efforts are in sync with other U.S. Government counter-terrorism activities and responses.

Content Production:

The GEC and its partners have established programming across multiple platforms, including social media, satellite television, radio, film, and print, in various languages.

Measurement & Evaluation:

The GEC was established with the aim of employing an agile response to terrorist communications, combining expertise from data science as well as the counter-terrorism field. According to its official website, the GEC 'approaches the task of undermining terrorist ideology with the understanding that the people and groups closest to the battlefield of narratives are the most effective in countering them'. The work of the GEC is therefore spread across four core areas: science and technology, interagency engagement, partner engagement, and content production. The incorporation of data science and technology expertise has enabled the development of good practices in measurement and evaluation of communication campaigns, built into a 'hypothesis-driven experimentation' approach that applies a "create-measure-learn" framework to activities to maximize effectiveness, including through A/B testing and multivariate analysis.[110] Furthermore, the 2010 National Framework for Strategic Communication report notes that all U.S. Government strategic communications program development 'should also include specific budgeting and resourcing for measurement activities that are needed to evaluate success'.[111]

Transparency & Acknowledgement of Risks and Challenges:

The National Framework for Strategic Communication of 2010 also details transparently the difficulties faced in measuring the success of communications-based responses in achieving attitudinal change: 'First, these efforts often target audiences' perceptions, which are not easily observed and, therefore, not easily measured... Second, it is difficult to isolate the effect of communication and engagement from other influences including other policy decisions. Lastly, communication and engagement effects are long-term and require persistent measurement'.[112] Because of these challenges, it is best to develop phased, layered plans for measuring success that are specific to a given plan or program.

110 *Ibid.*
111 The White House, *National framework for strategic communication*, 2010, p. 13.
112 *Ibid.*, p. 13.

Metrics

Governments are encouraged to develop, in collaboration with a range of stakeholders including the ICT industry, civil society organizations, and academic institutions, realistic indicators to measure the success of policies and programs aimed at preventing and countering violent extremism online. These indicators should be developed in line with human rights provisions, such as rights to freedom of expression, freedom of religion or belief, and the prohibition on arbitrary or unlawful interference with privacy, as enshrined in the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights.[113]

Indicators or metrics for the measurement of communications-based responses should be aligned with the objectives and the theory of change set at the start of the design process. Baselines and control groups should be used where possible to determine (positive or negative) changes in key metrics, and delineate the possible impact of a campaign on the target audience. These indicators can be broadly categorized into awareness, engagement and impact metrics, and can be combined and analyzed to build a comprehensive picture of a campaign's performance and impact.

Awareness, Engagement and Impact

Awareness metrics illustrate the reach of a campaign, or the number of people who are exposed to the campaign and their characteristics. Common awareness metrics for online content include impressions (the number of screens content appears on) and views (the number of people who actively consume content). Awareness metrics can also include demographic information, including the age, gender, and approximate location of audiences, as well as information related to the audiences' interests.

Engagement metrics illustrate the volume and types of interactions between audience members, campaigners or campaign content. Engagement metrics can include social media interactions such as likes, reactions, comments or shares, and can be positive or negative. The number and nature of engagements can help campaigners understand their audience's interactions with and reactions to a campaign or its content.

Impact metrics demonstrate a measurable change in the target audience's knowledge, attitudes or behavior that can be attributed to exposure to or engagement with campaign content. Awareness and engagement metrics, when properly analyzed, can be brought together to help evaluators understand the impact of their campaign. Additional indicators, such as evidence of offline action, responses to a call to action, or the qualitative evaluation of online comments, can contribute to the overall assessment of impact.

113 As UN Security Council Resolution 2354 (2017) notes, the right to freedom of expression is reflected in Article 19 of the Universal Declaration of Human Rights adopted by the General Assembly in 1948 (UDHR), and in Article 19 of the International Covenant on Civil and Political Rights adopted by the General Assembly in 1966 (ICCPR) and stresses that any restrictions thereon shall only be such as are provided by law and are necessary on the grounds set out in paragraph 3 of Article 19 of the ICCPR.

Monitoring & Evaluation Tools

The performance of online communication-based responses can be tracked using a variety of online analytics tools, including the "back-end" analytics provided on many social media platforms. These tools can provide a range of metrics and insights into the extent to which communications-based responses are reaching the intended audiences, and how these audiences are engaging with the content of campaigns. They can enable an iterative process allowing a campaign to be optimized and adapted to ensure it achieves is goals and objectives.

Communications-based responses to violent extremism and terrorism online, especially those conducted by civil society, remain in their relative infancy, and civil society organizations are often unfamiliar with best practices in online monitoring and evaluation. Governments can therefore encourage more sophisticated approaches by funding and supporting innovative data gathering, analysis and research methods in order to move beyond the basic analytics and metrics provided as standard on social media platforms.

There is a huge range of applicable analytics tools available, ranging from free open-source options through to more advanced commercial tools:

- **Social listening tools** can assist in the effective design and measurement of online communications-based responses. Such tools can identify public social media content across major social media platforms, such as Twitter, or forums and blogs such as Reddit or 4Chan. Content can be sorted relating to a topic, timeframe or language. The metrics provided by such tools can help to track narrative trends, uncover relationships between topics and reveal content, platforms, influencers, and language used by violent extremists or terrorists online, or those used by target audiences.
- **Network mapping** tools can help to visualize the online networks of violent extremist or terrorist groups, and the relationship of these groups with different audiences. Mapping tools can also help to understand the audiences that are interacting with communications-based responses and how campaign content is reaching certain audiences. Network analysis can also help to highlight online influencers who might provide exposure to target audiences for relevant campaigns.
- **Sentiment analysis** is the combined use of data mining and Natural Language Processing (NLP) to gather samples of text and analyze it for meaning using an automated process. Natural language processing software can be applied to samples of online text to classify, analyze and determine the meaning of large volumes words, phrases or sentences. This type of approach can help process data that is too large to analyze manually, deriving more in-depth quantitative insights from data that has been gathered throughout a campaign and helping to determine impact.

Qualitative Methods

Alongside the capabilities offered by online tools and analytics, there are a range of both online and offline qualitative approaches that can play an important role in the monitoring and evaluation of communications-based responses. These range from qualitative assessments of online engagements (e.g. comments) to offline surveys, focus groups, and interviews with

relevant target audiences. Qualitative approaches may be more expensive or time consuming than quantitative methods, but can provide valuable insights throughout a campaign. Such approaches are commonly employed in other fields to understand the resonance of communications, from social or political research to psychology, and best practices from these areas should be applied where appropriate.

Governments should be aware that such approaches may not be feasible with particular types of target audiences, particularly those with grievances towards the state, or that express sympathy for violent extremist or terrorist groups or narratives. When using in-person qualitative approaches, Governments should always operate transparently, consider who is best placed to facilitate or mediate, and allow participants to contribute anonymously when required to ensure an open and honest environment.

Monitoring & Evaluation Challenges

There are a number of inherent challenges in effectively monitoring and evaluating communications-based responses to violent extremism and terrorism online. For responses aimed at more downstream audiences, small sample sizes can limit the statistical significance of findings. Barriers to accessing certain audiences can also result in limited availability of the required metrics, and as a result an incomplete assessment of the impact of certain types of responses. This can result in a bias towards more easily accessible digital evaluation methods, resulting in a lack of qualitative data and nuance in the final assessment.

Even when qualitative methods are employed, there may be a “social desirability” incentive for participants to provide the outcomes that they believe evaluators are seeking, or that are considered socially acceptable. Careful evaluation design, and the effective implementation of suitable methods by appropriate actors, can reduce some of these potential effects. Comprehensive evaluation frameworks should therefore be transparent about the limitations of the methods employed, and any insurmountable limitations should be acknowledged in the final assessment. In order to avoid bias and provide an external, objective assessment, independent evaluations should be considered where appropriate.

Monitoring & Evaluation Risks

As well as the potential challenges of evaluating communications-based responses, these processes may carry ethical risks, for example through the inadvertent sharing or publication of identifiable online user data. It is therefore important to consider the legal context in which a campaign takes place, including privacy and data handling and protection laws. Governments should ensure that appropriate safeguards are in place for any Government-led responses, but also ensure that similar processes are in mandatory for all non-government responses that receive either Government funding or support.

In the interests of transparency, evaluations of communications-based responses should be shared with relevant stakeholders whenever possible to share learnings, improve the effectiveness of responses and build trust and credibility. However, there is a need to ensure that the privacy of those delivering the program, and the audiences it reaches, is protected by anonymizing any identifiable information. This could include user or account names, profile

pictures, or geo-location data. Any extracts of text produced by target audiences should also be modified sufficiently to prevent identification through social media search functions or search engines.

Case Studies: UK Government Communication Service Evaluation Framework & Guide to Government Communications Campaign Planning [114]

Common & Comprehensive National Framework:

The UK Government's 2016 Government Communication Service (GCS) Evaluation Framework is a tool available to stakeholders across the UK Government to help communicators measure and demonstrate the impact of Government communications work. The Framework is designed not only for public communications relevant to violent extremism and terrorism prevention, but activities directed towards a whole range of public service goals.

Learning from Other Sectors:

The Framework builds on latest industry standards and practices, learning from the private sector's experience in communications evaluation. This includes integration of methods to reflect a variety of communications mechanisms, including media and digital platforms, and consideration of the importance of measurement and evaluation from the outset of any communications effort.

Metrics:

The GCS Evaluation Framework encourages the use of 'a mix of qualitative & quantitative methods (e.g. surveys, interview feedback, focus groups, social media analytics, and tracking)' to measure the outcomes and impact of a communications campaign. The guide also suggests using 'benchmark' measurements to ensure a robust evaluation of change is acquired.[115]

Ongoing Measurement & Optimization:

The GCS framework includes suggestions for iterative adjustments to communications campaigns based on ongoing measurement and evaluation research. The framework suggests that users should: 'Review performance and ensure evaluation insights are fed into live activity and future planning'.[116]

114 Government Communication Service (GCS), *GCS Evaluation Framework*, January 2016; GCS, *A guide to campaign planning*.
115 GCS, *GCS Evaluation Framework*, p. 3.
116 *Ibid.*, p. 2.

Lifecycle Evaluation:

The UK Civil Service Government Communications Planning Guide is a tool to provide all UK Government employees with specific steps to be taken in the planning and deployment of any government-led communications, even before the production or dissemination of content starts. The guide includes steps for the identification of communications objectives, target audiences, content ideas, implementation mechanisms and evaluation processes: the 'OASIS key steps'.[117] It also provides links to tools that can help improve audience insights and the measurement of effectiveness of communications campaigns, such as links to social media analytics tools and guidelines.

C: Ethics & Security Risks

Multi-Stakeholder Approaches

Government communications may not be received as intended, and can reach a different audience than intended. Considering these risks, Government communications may be most effective in an upstream or preventative capacity, promoting social cohesion and resilience building. In these types of communication efforts, the potential consequences of the risks outlined above are less severe than in downstream communications. Governments can therefore work alongside the ICT industry and relevant civil society organizations, on a voluntary basis, to support and empower credible voices to ensure they are heard online, and provide both positive alternative messages for those vulnerable to violent extremist and terrorist content, and online engagement with individuals expressing violent extremist views or support for terrorism online.

Transparency

Where Governments directly conduct communications-based responses, it is important that these types of campaigns are transparent with regard to their origin or funding to avoid exacerbating grievances that violent extremist and terrorist groups exploit. Any messaging, either online and offline, must be complementary with broader Government policy and conduct in order to avoid undermining credibility.

A transparent approach can help to build trust between citizens and the State, reducing the risks outlined below. Where Governments support non-government or civil society organizations, transparency (in terms of both funding and support) is important to avoid undermining the credibility and impact of such responses, as well as encouraging the sharing of best practices and establishing a culture of learning and sharing among all stakeholders.

117 GCS, *A guide to campaign planning*, pp. 1–2.

Unintended Consequences

Government communications-based responses can have a range of complex impacts, not all of which will necessarily be positive. In the design process of any communications response, the potential positive impacts should be weighed against the potential negative or unintended impacts to understand their potential value, as well as undertaking efforts to mitigate any potential negative outcomes.

As a result, Governments can seek to work with a variety of stakeholders to identify where each actor can have the most impact, integrate offline and online activities, and adopt a "do no harm" approach with appropriate safeguards to ensure communications-based responses are proportionate and do not create unnecessary risks or unintended consequences.

These unintended consequences could include:

- The misunderstanding or trivialization of grievances in target audiences;
- The reinforcement of the appeal of violent extremist or terrorist narratives;
- The risk of stigmatizing certain groups of citizens as 'at risk', or further alienating or excluding certain groups that are distrustful of the State;
- The undermining of campaigns' legitimacy or credibility in challenging violent extremist or terrorist narratives through affiliation with brands or messengers that may lack credibility in their target audiences.

Security Risks

Communications-based responses to violent extremist and terrorist content can put both audiences and campaigners at-risk, potentially exposing participants to online abuse or, in extreme cases, physical harm. Using a "do no harm" approach, the safety of the individuals carrying out communications-based responses should be of paramount importance, and those involved should be provided with a thorough assessment of the potential risks.

This assessment should form the basis of a security and ethics framework, including steps to mitigate these risks. This framework should be agreed by all stakeholders during the design phase of any response and should also be regularly consulted and updated as required during the delivery and monitoring and evaluation phases.

In some very rare cases, governments chose not to reveal their support for certain communications-based responses due to security concerns, for example a "downstream" response (*see Chapter 5: Figure 1*) where the audience of a campaign may target or threaten those delivering it. Such concerns may also pose an ethical risk, in the form of exposing non-government stakeholders to heightened security risks.

The specific security and ethical risks that must be considered will vary depending on the type of campaign, as well as other factors such as the contexts in which it is delivered, and the intended target audiences. However, these risks can be effectively mitigated through careful planning and implementation, including more granular targeting and careful content choices.

Engagement Risks

Finally, whilst many communications-based responses do not seek direct engagements with individuals who are at-risk or vulnerable to violent extremism or terrorism, or who are currently members of violent extremist and terrorist groups, all responses should have pre-determined guidelines in place for any interactions that do occur with such individuals. This should include guidelines for both the delivery and evaluation phases of programs (e.g. focus groups, surveys).

Acting in an appropriate fashion when interacting with a potentially vulnerable individual is not only an ethical requirement, but also a potential opportunity to achieve a positive impact for certain types of responses to counter violent extremist and terrorist recruitment online. Relevant considerations might include:

- How to respond to vulnerable individuals in a way that reduces their personal risk and sensitively and effectively addresses their specific needs;
- How to avoid overreaching into activities that individual campaigners are not qualified to undertake;
- Which are the appropriate authorities, civil society or community support options to connect vulnerable people with if required.

Identifying and responding to potentially negative, dangerous, unexpected or counter-productive reactions to communications campaigns can be aided by the work of a campaign manager. Campaign managers are often used in communications campaigns unrelated to terrorist or violent extremist content, such as commercial advertising, and can be helpful in both identifying and, where appropriate, responding to comments, reactions or activity connected to communications content.

Additionally, most large social media platforms provide advertisers or users with the ability to view, analyze and respond to comments and reactions to posted content. There are also a host of commercial tools available to streamline campaign management for teams producing or analyzing the impact of a large number of online communications campaigns simultaneously.

Finally, transparency is key in engagement moderation approaches. Government agencies should consider publishing a social media policy on any public communications pages or sites, which outlines the guidelines for content that could be subject to moderation by campaign managers. This might include threats or violent content, for example.

Communications-Based Responses:

5. Collaboration with ICT Industry and Engagement with CSOs

This Chapter seeks to provide policymakers with practices and case studies to spur effective collaborations between Governments, the private sector and civil society, where appropriate, to design and deliver a range of effective communications-based responses to address all aspects the threat from violent extremism and terrorism online. This Chapter is divided into two sub-sections: Government Partnerships with ICT Industry & Civil Society; and Partnerships Across a Spectrum of Communications-based Responses.

Relevant Good Practices from the London-Zurich Recommendations:

Good Practice 6: *To adopt a multi-stakeholder approach between Governments, the ICT industry and civil society organizations in preventing and countering violent extremism and terrorism online.*

Good Practice 13: *To address all aspects of violent extremism and terrorism by tailoring online interventions to take into account a spectrum of communications responses, including preventative programs and counter-narrative campaigns.*

Good Practice 14: *To encourage voluntary collaboration to produce authentic and innovative communications-based approaches to the challenge of violent extremist and terrorist content online by convening the ICT industry, civil society organizations and other actors.*

Good Practice 15: *To ensure campaigns have a distinct target audience (or audiences), a specific goal (e.g., to decrease the risk of radicalization to violence or promote peaceful alternatives to violent narratives) and provide tightly focused, distinct, and context-specific messages. Analysis of specific audience(s) can enable the identification of suitable messengers that are credible to the relevant target audience(s).*

INTRODUCTION

Given the transnational nature of the Internet, preventing and countering violent extremism and terrorism online can be supported through effective collaboration between Governments and a variety of stakeholders, including the ICT industry and relevant civil society organizations. The Zurich-London Recommendations emphasize that: "States have the primary responsibility in countering violent extremism and terrorism. It is a State's prerogative to decide which approach is most effective, in compliance with its obligations under international law as well as in accordance with its national law."[118]

A multi-stakeholder approach combining political, technical, and contextual expertise, and based on up-to-date and rigorous research on the nature of drivers to support for violent extremism or terrorism can play an important role in effective communications-based responses. Such collaborations help to harness the necessary creativity, expertise and resources, and encourage the development of innovative and sustainable responses with clear strategies and effective planning, design, delivery and evaluation.

A: Government Partnerships with ICT Industry & Civil Society

Multi-Stakeholder Approaches: Civil Society

Governments considering communications strategies to challenge violent extremist or terrorist content online should recognize the contribution civil society and other civic organizations can provide as implementers, managers or creators of campaigns, and not solely as partners or constituencies involved in their dissemination. Civil society organizations, which are typically deeply engaged with local communities, are more likely to be credible among key audiences and can be effective partners in building impactful and sustainable local community-level responses.

Civil society organizations can provide authentic voices for communications-based responses to a range of target audiences, including from specific gender or age perspectives, or from community groups such as faith organizations or education institutions. Partnerships with civil society organizations can therefore ensure that communications-based responses take into account important dimensions of the violent extremist or terrorist recruitment dynamic, such as gender, and that they address the specific concerns and vulnerabilities of relevant target audiences.

Multi-stakeholder approaches are most likely to be effective when all participants share a common understanding of each other's respective roles, responsibilities, strengths and limitations in responding to violent extremism and terrorism online. Given the potential limitations of State involvement in responding to these challenges (*see Chapter 4: C: Ethics and Security Risks in Communications-Based Responses*), Governments can encourage a wide range of civil society groups to contribute to building reliance among communities and push

118 GCTF, *Zurich-London Recommendations on Preventing and Countering Violent Extremism and Terrorism Online*, 2017, p. 4.

back against the radicalization and recruitment narratives and tactics of violent extremist and terrorist groups.

Due to the breadth of potential approaches to communications-based responses, Governments can seek to build partnerships with civil society organizations beyond the P/CVE sector. This could include organizations whose primary focus is promoting human rights, working with youth, providing social services and support, or delivering cultural activities. Such organizations may not have considered P/CVE as a part of their mandate, or be aware of the vital role that such approaches can play in a broad, whole-of-society approach to preventing and countering violent extremism and terrorism online.

Governments can therefore support and build the capacities of civil society actors by providing training, resources (e.g. toolkits), and/or funding to encourage greater participation in P/CVE from established organizations both within and outside the sector. These efforts must be long-term and sustainable as civic efforts will invariably improve over time. Initial programs, especially from organizations new to P/CVE may take time to evolve and demonstrate their impact. Governments should therefore also invest in the monitoring and evaluation of both their own capacity building efforts, and the civil society programs they support.

Building Trust

An open and honest dialogue between all stakeholders is vital to pursue efficient and sustainable collaboration to prevent and counter violent extremism and terrorism online. Governments should be aware of any existing sensitivities and avoid securitizing multi-stakeholder relationships, especially in instances where civil society may be concerned about the stigmatization of particular communities.

Civil society organizations' involvement in counter-messaging with Governments should therefore be voluntary and based on trust, confidentiality, incremental buy-in and commitment. Discussions about the potential backlash against civil society organizations that receive direct funding or commissions for communications work from Governments should be open and transparent in any development of new partnerships. This allows civil society organizations to make informed decisions on whether to collaborate with Governments on responses to the challenges of violent extremism and terrorism online.

Case Study: Building a Stronger Britain Together

The Building a Stronger Britain Together (BSBT) program supports civil society and community organizations in the UK that aim to create communities resilient to extremism and that seek to produce positive alternatives to extremist recruitment. The program, funded by the UK Home Office and managed by UK Community Foundations and private sector communications agency M&C Saatchi, allows community organizations to bid for in-kind support or grant funding for programs that attempt to deliver objectives relevant to the UK Government's CONTEST goals

in local communities.[119] The program therefore sits as part of a national strategy for prevention and countering of violent extremism and terrorism, as encouraged in Chapter 4. As of February 2019, 233 community organizations had been successfully awarded grants or in-kind support through the BSBT program.[120] Reported outputs as of summer 2017 included: 20 communications strategy packages, 15 website builds, 33 training packages and 5 social media campaigns.[121]

Whole-of-Society Approach:

BSBT is broadly aimed at combating 'extremism in all its forms' and supports a wide range of community organizations, 'regardless of race, faith, sexuality, age and gender'.[122] The organizations partnered with BSBT include community centers, faith, cultural and youth groups, and sports programs. BSBT also recognizes that the use of the internet for violent extremist and terrorist purposes is a continuing trend and therefore encourages applications that seek to 'to promote positive alternative narratives to counter extremist content online and/or challenge extremist activity online'.[123]

Support for this wide range of community organizations is based in the recognition that local organizations and civil society actors "have an unrivalled understanding of local needs and challenges and are best placed to deliver grants to local organizations".[124] Projects are varied in their target audiences, and include community support for vulnerable and isolated women in ethnic minority communities, and workshops addressing British values and extremism in local contexts.[125] This inclusive approach helps to ensure that projects reach at-risk and marginalized populations through credible voices, outside of the UK Government.

BSBT also seeks to strengthen the relationships between member organizations and facilitate the sharing of 'good practice' by hosting regional events that also offer training in areas such as the use of social media and effective public relations.[126] The BSBT fund is managed by private communications company M&C Saatchi, which ensures that expertise from the commercial sector is present alongside Government and civil society voices.

119 Home Office, *Guidance Building a Stronger Britain Together*, 16 September 2016.
120 See https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/777653/Building_a_Stronger_Britain_Together_partners.pdf.
121 Home Office, *Partnership Support Programme Summer 2017 Update*.
122 Home Office, *Building a Stronger Britain Together*.
123 See https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/759649/bsbt-inkind-guidance-applicants.pdf, p. 1.
124 https://www.efc.be/news/new-programmeme-building-stronger-britain-together-deliver-800000-grants/.
125 Home Office, Partnership Support Programme Summer 2017 Update.
126 *Ibid*.

Clear Strategy:

BSBT supports a broad spectrum of organizations working to uphold the British values of "democracy, free speech, mutual respect and opportunity for all". Within this remit, BSBT's application process ensures that successful applicants specifically align with and uphold the UK Government's existing counter-extremism strategy.[127] Prospective BSBT organizations are asked to assess their project's relevancy to four desired outcomes:

1. Fewer people holding attitudes, beliefs and feelings that oppose shared values;
2. An increased sense of belonging and civic participation at the local level;
3. More resilient communities;
4. Targeted activity to counter known extremist activity at local level.[128]

In order to be considered for a grant or in-kind funding, proposals must align with outcome four and at least one of the other three outcomes.[129]

Measurement & Evaluation:

Applications for BSBT support also ensure that projects are set up with a designated target audience and measurable outcomes in place, including processes for how these outcomes will be recognized and measured through key performance indicators (KPIs).[130] Organizations must identify their broad aims and objectives, as well as those of the specific project that they wish to be supported through BSBT and how these relate to the four BSBT outcomes.

Transparency:

Successful applicants must openly state that they are the recipient of government funding on their website/other relevant communication, or else have their funding rescinded.[131] In addition, a list of funded organizations is published annually and made available to the public.

Proactive & Voluntary Cooperation with ICT Companies

Voluntary and transparent collaboration between Governments and ICT companies can help to achieve a better understanding of the threat of violent extremist and terrorist communications online, and enhance the impact of communications-based responses in challenging these threats. As in partnerships with civil society organizations, Government partnerships with ICT companies regarding communications-based responses should be based on transparency and trust and in accordance with national legislation.

127 Home Office, *Applying for grant support Guidance for applicants*, 2019.
128 *Ibid.*, pp. 13-15.
129 *Ibid.*, p. 13.
130 *Ibid.*, p. 16.
131 *Ibid.*, p. 7.

Governments can encourage ICT companies to be proactive in preventing and countering violent extremism and terrorism on their platforms, and to better protect their users by supporting innovative communications-based approaches from civil society. Such collaborations with major ICT companies can help to enhance the reach and impact required to effectively and sustainably counter the threat posed by violent extremist and terrorist communications online.

Examples of such collaboration include a partnership between UNDP and YouTube's *Creators for Change* program that supports local CSOs in Malaysia, Indonesia, Thailand and the Philippines through small grants, mentoring and the development of a network to encourage youth influencers to create content that counters terrorist and violent extremist messaging and offers a positive alternative online.[132] UNDP has also built a similar partnership with Facebook in the region to deliver a series of online videos featuring former violent extremists.[133] On the national level, the Australian Government partnered with a wide range of private sector technology companies, including Facebook, Google, Microsoft (including Xbox), Oath, Twitter, Instagram, Periscope and Yahoo to host DIGI Engage 2018 in conjunction with the ASEAN-Australia Special Summit.[134] The event brought together 80 youth leaders from the region to enhance their skills and provide tools to help them contribute to countering violent extremism online. This initiative has been running annually since 2017, and continues to engage young people in the Asia-Pacific region.[135]

Governments should also seek to engage the hundreds of smaller ICT companies, beyond the largest social media platforms such as Facebook, YouTube or Twitter, that can also play a key role in the online violent extremist or terrorist ecosystem. The range of platforms used by violent extremist and terrorist groups typically varies by geographical context and language, but may include smaller social media platforms, forums, messaging services, or audio or video-based platforms. Governments should recognize that while smaller platforms may wish to contribute to preventing and countering violent extremism and terrorism online, they may not have the same level of resources or in-house expertise as larger companies, and may therefore require additional support from other members of the ICT sector.

International Coordination

Governments can work through existing Government or industry-led forums to share best practice and resources where possible, such as the United Nations Counter-Terrorism Executive Directorate (UN CTED), the Global Internet Forum to Counter-Terrorism (GIFCT) or the EU Internet Forum. Such platforms can help to develop shared objectives and frameworks, open communication channels, build capacity, defuse conflicts of interest and identify critical gaps. Coordinated efforts help to streamline multi-stakeholder initiatives and ensure complementary actions.

132 See http://www.asia-pacific.undp.org/content/rbap/en/home/operations/projects/overview/creators-for-change0.html.
133 See http://www.asia-pacific.undp.org/content/rbap/en/home/programmes-and-initiatives/extremelives.html.
134 See https://digiengage2018.splashthat.com/.
135 See https://digiengage2019.splashthat.com/.

Case Study: European Union Internet Forum

Voluntary Collaboration:

The EU Internet Forum was launched in December 2015 by the Commissioner for Migration, Home Affairs and Citizenship to address the exploitation of the internet by terrorist groups. The EU Internet Forum brings together EU Home Affairs Ministers, the internet industry and other stakeholders (such as Europol, the European Strategic Communications Network and the Radicalisation Awareness Network) to work together in a voluntary partnership. The aim of the Forum is to address this issue of terrorists' use of the internet and therefore to better protect EU citizens. As such, the EU Internet Forum has two key objectives: to reduce availability of and accessibility to terrorist content online; and to empower civil society partners to increase the volume of effective alternative narratives online. The EU Internet Forum provides a chance for representatives from EU Member States to discuss issues concerning terrorists' and violent extremists' use of the internet with a broad set of ICT industry companies, in December 2018 including representatives from Baaz, Dropbox, Facebook, Google, Justpaste.it, Microsoft, Snap and Twitter.

Supporting Civil Society Communications-Based Responses:

In December 2016, the EU Internet Forum launched the EU Civil Society Empowerment Programme to help develop alternative and counter narrative campaigns online.[136] The program now provides nearly €12 million of EU funding to civil society groups to support the development of campaigns to challenge extremism and terrorism. The program acknowledges the fact that many civil society organizations already actively attempt to provide alternative narratives to violent extremist and terrorist narratives, but that they often lack capacity and resource to do so effectively online. The program seeks partnerships with marketing and communications experts and creatives to provide training to civil society grantees, as well as its existing partnerships with large social media companies, who also provide training and best practice from the world of online marketing and content creation. The training resources from the Civil Society Empowerment Programme area available online in a number of languages.[137]

Transparency:

Civil society organizations that receive funding and support from the Civil Society Empowerment Programme are listed in a public database online, along with the project title and the amount of funding granted.[138]

136 See https://ec.europa.eu/home-affairs/what-we-do/networks/radicalisation_awareness_network/civil-society-empowerment-programme_en.
137 *Ibid.*
138 See https://ec.europa.eu/home-affairs/sites/homeaffairs/files/financing/fundings/security-and-safeguarding-liberties/internal-security-fund-police/union-actions/docs/isfp-list-proposals-selected-for-funding-during-2018_en.pdf.

Private Sector Expertise

The ICT sector can offer significant technical expertise, training and resources to both Governments and civil society to support the successful implementation of communications-based responses. This includes advice on how best to reach and engage specific audiences on specific ICT platforms (including through targeted advertising), and how to effectively assess the impact of online communications. ICT companies can also play an important role in funding and supporting research into the use of their platforms by violent extremist and terrorist groups. This research can then be utilized by partners across different sectors when designing communications responses that are tailored to the online ecosystem.

In working with the private sector to tackle violent extremist or terrorist communications online, Governments should also seek to learn from and collaborate with a broad range of experts from other fields, including data analytics, online communications, advertising, marketing and content production. Expertise from these areas can help to move P/CVE responses beyond basic content marketing approaches towards more sophisticated online campaigns, in line with the latest approaches employed in the commercial sector.

For example, many commercial brands are moving away from a focus on specific products towards more values-based campaigns in order to capture the interest of younger audiences. Others are creating immersive campaigns with multiple versions of content to enable audiences to find the messaging that most resonates. Alternatively, as the value of big campaigns has faded, many companies are increasingly utilizing micro-targeted campaigns featuring "influencers" to promote their products to highly specific audiences. These approaches have been developed through in-depth market research and the application of online analytics to build a clear understanding of the audiences that they are intended to reach. In some instances, campaigns and content are starting to be developed and iterated through artificial intelligence.

Government and civil society communications are often late to adopt these more cutting-edge approaches, instead favoring broad campaigns for mass audiences. As expertise in these areas can be prohibitively expensive for civil society organizations, Governments can also contribute by encouraging more pro-bono or in-kind support from the private sector through Corporate Social Responsibility (CSR) schemes to address the challenges of violent extremism and terrorism online.

B: Partnerships for a Spectrum of Communications-Based Responses

There is often a lack of conceptual clarity surrounding the spectrum of possible communications-based responses to prevent and counter violent extremism and terrorism online. In some instances, "counter-narratives" is used as a catch-all term to describe a range of approaches to messaging in P/CVE practice. It is therefore vital to accurately and consistently differentiate between the different types of communications-based responses to ensure they are employed correctly, and are not targeted at inappropriate audiences to avoid generating unintended consequences. For the purposes of this toolkit, communications-based responses are broadly divided into 'upstream' and 'downstream' approaches (*see Chapter 4: Addressing All Forms of Violent Extremism and Terrorism*).

It should be noted that, despite these categorizations, the range of communications-based responses are not discrete and operate along a spectrum, and as a result, some campaigns or programs may incorporate elements of one or more types of response. Additionally, communications-based responses are highly context-specific, so may not be directly comparable across different national or local settings. The table below (*Figure 1*) outlines the differences between upstream and downstream approaches to communications-based campaigns, including the variation in objectives, types of messenger required, and types of audience targeted.

Figure 1: Upstream and Downstream Approaches to Communications-Based Responses

	Upstream Approaches		**Downstream Approaches**	
Type of Response	Public awareness	Positive or alternative narratives	Counter-narratives	Online engagement and interventions
Goals	Prevention & Resilience Building		Deter engagement with violent extremist or terrorist content, or reverse early stages of radicalization	Encourage deradicalization or disengagement from violent extremist or terrorist groups or ideologies
Objectives	• Communicate Government policy, strategy or legislation • Raise awareness of support services • Refute disinformation or misinformation • Address concerns and build public trust and relationships with key constituencies • Build a strong and inclusive sense of identity and belonging • Raise awareness of citizen rights and responsibilities	• Promoting positive values e.g. human rights, democracy, tolerance, diversity, pluralism • Promoting pro-social avenues for citizen participation • Challenging negative stereotypes or prejudices	• Directly challenge, deconstruct, discredit or demystify violent extremist or terrorist messaging and ideologies through emotive content, exposure of hypocrisy, lies and disinformation or misinformation	• Directly engage members of violent extremist or terrorist groups or online communities

	Upstream Approaches		Downstream Approaches	
Messengers	• Government • Public officials • Politicians	• Civil society • Communities • Youth • "Influencers" e.g. figures from sport or entertainment • Private sector • Religious figures or institutions • Former extremists / terrorists and/ or survivors of extremist or terrorist violence		• Trained interventions practitioners and/or former extremists / terrorists
Target Audiences	Broader audiences e.g. general public, parents and families, public officials, practitioners, secondary age youth		At-risk or vulnerable audiences e.g. those actively watching or searching for violent extremist or terrorist content online	Members of violent extremist or terrorist groups or online communities

Governments should be aware of the potential to undermine a campaigns' credibility with certain target audiences when directly delivering, participating in or openly endorsing downstream communications-based responses. Instead, Governments can help to build the capacity of organizations working in these areas, and encourage others (such as the private sector or civil society foundations) to provide direct support.

Case Study: UNESCO Prevention of Extremism through Youth Empowerment:

On February 1, 2018, UNESCO launched a $2 million two-year project to engage youth in Tunisia, Morocco, Libya, and Jordan in the prevention of violent extremism. UNESCO and the United Nations Counter-Terrorism Centre (UNCCT) launched the project, "Prevention of Violent Extremism through Youth Empowerment in Jordan, Libya, Morocco and Tunisia", with an event at UNESCO Headquarters in Paris in April 2018 and with co-funding from the Canadian Government. The program is aimed at building youth empowerment to foster resilience to violent extremism, rather than direct counter-messaging to specific extremist narratives and content.

Empowering Youth as Credible Messengers for Alternative Narratives:

The UNESCO program focuses on the role of youth in responding to violent extremism in the region. The launch included the participation and remarks of six young people affected in some way by violent extremism in the region. The project supports youth-driven initiatives in education, sciences, culture, and the media to prevent violent

extremism. Youth organizations, education stakeholders and media professionals are involved in the multi-stakeholder project, working together across a variety of areas that include youth dialogues, training in conflict-sensitive reporting and critical thinking labs.[139]

The project provides training on "countering online hate speech" and seeks to develop "new media spaces to disseminate alternative narratives by and for youth", among other elements. To this end, it mobilizes media professionals and online youth communities through training sessions and the development of national and regional online campaigns.[140] UNESCO states its aim to "create opportunities for young women and men to engage as change-makers and peacebuilders in their immediate communities and wider societies, and to promote a constructive vision of young people as leaders, addressing hate related issues".[141] In line with UN Security Council Resolution 2250 and the UN 2030 Agenda for Sustainable Development, the project aims to build skills that can be used online and offline.

Collaboration with Multiple Stakeholders:

UNESCO work closely with partners such as Ministries of Youth, Education, Labor and ICT companies, as well as with civil society organizations that include and reach out to youth, educational and cultural networks, local religious leaders and universities. Across all of these partnerships, UNESCO enshrines the principles of human rights and transparency.

In an event of November, 2018 held in Canada, the EU and UNESCO hosted a youth seminar on media, journalism and culture for human rights through the program. Participants included civil society organizations, students of journalism and media. Three sessions were held on the discussion topics of media information literacy, human rights sensitive reporting, and cartoons as a cultural tool for building tolerance and openness.[142]

Messages, Audiences & Messengers

The messages in communications-based responses should be tightly focused, distinct, and context-specific, with a focus on creating compelling narratives and content. Campaigns designed to build resilience to violent extremism or terrorist content will require a different message and messenger for impact from campaigns designed to disengage or deracialize supporters or sympathizers of violent extremist or terrorist groups.

139 UNESCO, *Launch of Project to Tackle Violent Extremism in Jordan, Libya, Morocco and Tunisia*, 19 April 2018.
140 UNESCO, *New project to tackle violent extremism in Jordan, Libya, Morocco and Tunisia*, 05 February 2018.
141 See https://en.unesco.org/preventing-violent-extremism/youth/project.
142 UNESCO, *Canada, EU and UNESCO host media, journalism and culture for human rights youth seminar*, 18 November 2018.

If campaigns are not designed coherently, there is a possibility of unintended consequences, including inadvertently aiding the dissemination of or support for violent extremist or terrorist narratives for more downstream campaigns.[143] When designing such messages, a key consideration is that efforts to debunk or disprove violent extremist or terrorist messaging can instead embed the narratives more deeply within the target audience.[144] Exposure to information that challenges an individuals' views can serve to entrench existing views.[145] The message should therefore be planned with, and tailored to, the specific audience in mind, and with an awareness of the potential effects on those not within the target audience that may still be exposed to the campaign.

Messengers must be authentic and credible to the target audience in order for the message to resonate effectively, and may include a member of the target audience themselves. Gender should also be an important consideration, as campaigns may resonate differently with men and women, or boys and girls.

Some of the most effective campaigns directly engage or involve audiences through focus groups or surveys during their development phase to test, shape and refine messages, content or dissemination plans, as they often have relevant direct experiences, local knowledge, and an understanding of how best to engage and influence their peers. The ethics, security and risk considerations of such engagement need to be considered in advance, including issues such as incentives, anonymity, and how results are acquired and recorded. Necessary steps must be taken to protect personal details and identities of all participants.

Case Study: Afrika Moja – Intergovernmental Authority on Development (IGAD) Center of Excellence in Preventing and Countering Violent Extremism

In September 2017, IGAD convened young people from across the Horn and Eastern African region, including Somalia, Djibouti, Kenya, Tanzania and Uganda, to design and launch a civil society platform to promote a series of alternative and counter-narrative campaigns based on stories from the region.[146]

The two-day workshop built on a previous event held to train youth activists with the skills required to deliver innovative and effective video and image-based campaigns,

143 Nicholas J. Cull, *Counter Propaganda: Cases from US Public Diplomacy and beyond*, Legatum Institute Transitions Forum, July 2015.

144 C.R. Sunstein, *On Rumors: How Falsehoods Spread, Why We Believe Them, and What Can Be Done* (Princeton: Princeton University Press, 2014), pp. 47-53. Another recent study concluded that during the 2012 US Presidential campaign "(...) Twitter helped rumor spreaders circulate false information within homophilous follower networks, but seldom functioned as a self-correcting marketplace of ideas." Cf. J. Shin, L. Jian, K. Driscoll, F. Bar, *Political rumoring on Twitter during the 2012 US presidential election: Rumor diffusion and correction*, New Media & Society, 08 March 2016, p. 2, doi: 10.1177/1461444816634054 (2016-10-23).

145 Amanda Ripley, *Complicating the Narratives – The Whole Story*, The Whole Story, 27 June 2018. Retrieved 18 September 2018 from https://thewholestory.solutionsjournalism.org/complicating-the-narratives-b91ea06ddf63.

146 IGAD Center of Excellence in Preventing and Countering Violent Extremism (ICEPCVE), *IGAD Launches Afrika Moja, An Umbrella Platform For Civil Society Campaigns To Counter Violent Extremism*, 20 September 2017.

using a peer-to-peer approach to ensures the target audience was involved in the creation of the content. This resulted in the creation of the Afrika Moja platform, which aims to challenge violent extremist messaging in the region by amplifying positive local stories and highlighting the hypocrisy of violent extremist groups. The platforms' first campaign, 'Strength in Diversity' was also created during the event, and aimed to emphasize the common values held across the continent. Following the workshop, the campaign was disseminated via social media (Facebook, Twitter and Instagram).The Center has continued to support CSOs and youth in the region to create effective counter-narrative campaigns, including through additional workshops in Uganda, Kenya, Ethiopia and Djibouti focusing on the amplification of young leaders' voices, the creation of counter-narrative illustrations and infographics, building effective strategic communications partnerships, and the effective use of video respectively.[147]

Alternatively, campaigns can leverage existing "influencers" – those with the ability to effectively reach and resonate with certain audiences – in both the design and delivery of campaigns. For example, depending on the audience, this could include figures from the local community or cultural figures (e.g. from music or sports).

Campaigns should therefore be informed by thorough target audience analysis, an offline understanding of the desired audience, and A/B testing of messages to determine effective and creative content approaches. This evidence-base should be utilized during the creative planning, content development and testing phases, and incorporated into detailed content plans to ensure campaigns are designed and delivered effectively, and meet their goals and objectives.

Communications Mediums

The medium as well as the content and delivery of messaging is important in ensuring the effectiveness of communications campaigns. This includes both the type of content (e.g. text, audio, video etc.) and the communications channel or platform through which it is disseminated (e.g. social media, gaming platforms, broadcast, print etc.).

Campaigns should be designed based on a strong awareness of relevant trends, including what is popular among the target audience and why. In some cases, this may not prove to be an online platform or type of content. Instead, an offline approaches and existing content (e.g. popular television or radio programs or print publications) may prove more influential. These considerations should inform detailed distribution plans for each campaign.

147 ICEPCVE, *Yali Workshop – 13th To 16th November 2018: Amplifying The Voices Of Young African Leaders*, 02 October 2018; ICEPCVE, *Using Illustrations And Infographics To Communicate Violent Extremism*, 07 July 2019; ICEPCVE, *Partnerships To Strenghten Strategic Communications*, 30 May 2019; ICEPCVE, *Tell Me A Story: Video Messages To Challenge And Undermine Violent Extremist Ideologies*, 21 May 2019.

Case Study: Duta Damai – Ambassadors for Peace

The Deputy for Prevention, Protection and Deradicalization of The Indonesian National Agency for Combating Terrorism (BNPT) established Pusat Media Damai (Peace Media Center) to support national efforts to counter and challenge violent extremists' and terrorists' use of the Internet. As part of its wider strategy, BNPT subsequently created Duta Damai – the Peace Ambassador Initiative - in 2016.[148]

Working with Youth:

Duta Damai is a community of youth bloggers, website creators, and designers that work under the overall strategy of the National Counter Terrorism Agency, helping to support the Government's counter-terrorism goals and helping to promote digital literacy, democracy and peace.[149]

The Ambassador programme teaches its youth membership to create and disseminate their own counter and positive narratives online and offline. These youth Ambassadors share their content with each other and with the public online through an array of different sources. Since its creation, the Peace Ambassadors programme has spread to 13 provinces and reached 780 youth.[150]

The programme considers sustainability and scale through granting the youth Ambassadors them the opportunity to become trainers themselves, helping other individuals in their communities to learn similar skills.

International Collaboration:

The Peace Ambassador Initiative has been extended into more provinces each year and this year was expanded globally to include other youth from Malaysia, Singapore, Cambodia, Laos, the Philippines, Brunei Darussalam, Myanmar and Thailand.[151] A three-day conference for the initiative's international work was hosted, with the objective of "Spreading Peace in Cyberspace". This convening gathered 116 young people between the ages of 20 and 30 from Indonesia and other ASEAN countries.

The first aim of this conference was to educate participants about the dangers of online radicalization and the ways in which extremist and terrorist narratives are spread over the Internet. The end goal was to empower youth to use cyberspace to challenge and combat these narratives with positive and peaceful ones of their own. The programme offered attendees both writing and technical skills (website/graphic design, video editing, etc.).[152]

148 Asmak Abdurrahman, *ASEAN Youth Ambassador for Peace 2019 Resmi dibuka*, Duta Damai NTB, 2019.
149 See https://dutadamainusatenggarabarat.id/tentang-kami-2/.
150 Dyah Dwi Astuti, *Duta Damai Dunia Maya direncanakan diperluas hingga antarbenua*, ANTARANEWS.com, 24 April 2019.
151 Asmak Abdurrahman, *ASEAN Youth Ambassador for Peace 2019 Resmi dibuka*.
152 *Ibid.*

Communications-Based Responses:

6. Empowering Youth and Building Resilience through P/CVE, Online Safety, and Digital Citizenship Education

This Chapter aims to provide policymakers and practitioners with practices and case studies on the role of education in communications-based responses to violent extremism and terrorism online. It outlines the roles of Government and other stakeholders, including the education sector, civil society and the private sector, and the range of possible approaches that can be employed to protect young people online. This chapter is divided into three sub-sections: Policy Design for Educational Responses; The Spectrum of Educational Responses; and Implementing Educational Responses.

Relevant Good Practices from the London-Zurich Recommendations:

Good Practice 3: *To develop a clear strategy to tackle violent extremism and terrorism online based on a whole-of-government and a whole-of-society approach, which coordinates both content- and communications-based responses, as well as offline activities, including education and engagement of civil society organizations where appropriate.*

Good Practice 14: *To encourage voluntary collaboration to produce authentic and innovative communications-based approaches to the challenge of violent extremist and terrorist content online by convening the ICT industry, civil society organization and other actors.*

INTRODUCTION

Education can play a crucial role as part of a broader communications-based strategy for preventing and countering violent extremism and terrorism online. Educational systems reach huge numbers of young people, who are often those most targeted by violent extremist and terrorist groups, but also provide a vast resource in terms of existing skills, approaches, networks and infrastructure. Education is vital to instill the positive values and skills required for young people to succeed in the digital age, and can spur positive societal change by encouraging young people to be active and engaged citizens online. Additionally, whilst the majority of adults do not participate directly in the formal education system, it can offer indirect channels to reach adults as well as youth, for example through schools' engagements with parents.

This chapter primarily focuses on the adoption and implementation of P/CVE policies and programs at the primary and secondary school levels, but also provides case studies from the informal education sector. Approaches that do not explicitly seek to address violent extremism or terrorism online, but that promote broader online safety and digital citizenship, are also included. As with all other forms of communications-based responses, educational approaches should be considered in the context for which they were designed and, where appropriate, should be adapted accordingly to address the specific local drivers of online violent extremist or terrorist radicalization and recruitment. Similarly, the specific educational context and system in which they were originally employed must be considered, given that these conditions can vary enormously between or even within countries.

A: Policy Design for Educational Responses

Whole-of-Society Approaches in Education

As with communications-based responses in general, Governments should promote a whole-of-society approach to educational responses, bringing together all relevant stakeholders and ensuring that efforts are complementary to a broader national strategy to prevent and counter violent extremism and terrorism online. Governments, the education sector, civil society, communities and families, and the private sector should work together to identify how education can be used effectively to build resilience and reduce recruitment and radicalization to violent extremism and terrorism. The impact of gender should be considered, with the design of policies and programs, where appropriate, taking into account the potentially different needs of young women and young men. Governments can play an important role in encouraging and supporting collaboration between educational institutions and this wide variety of stakeholders to create effective and sustainable educational responses, from initial convening and engagement to conducting needs-assessments, and designing, implementing and evaluating programs.

Given the potential unintended consequences of addressing sensitive topics in an educational setting, sufficient care should also be taken to ensure that the education sector is not overly securitized for educational approaches to be effective. Governments should ensure the appropriate use of terminology, and that educational initiatives to prevent and counter violent

extremism and terrorism online are framed and explained clearly to secure buy-in from all relevant stakeholders, as well as young people. It is therefore vitally important that all educational responses and approaches are fully transparent in terms of their origin, funding, and purpose to secure buy-in and avoid exacerbating any existing grievances that violent extremist and terrorist groups exploit.

Primary, Secondary & Tertiary Education

All levels of the education system can play a role in instilling positive values, building skills and resilience, and ultimately preventing and countering the effects of violent extremism and terrorism online. Many cognitive skills related to the formation of values and development of critical thinking are developed in early childhood. Therefore in primary education settings, the focus should be on more implicit approaches, including building positive values such as diversity and tolerance towards the attitudes of others, and developing basic online safety and early critical thinking skills. The consultation and involvement of parents and family members is especially relevant in primary settings given the sensitive topics such approaches can address.

At the secondary level, both implicit and explicit approaches can be effective, from building on the online safety and values-based approaches employed at the primary level, to a greater focus on positive online behaviors and active online citizenship approaches, as well as more directly addressing sensitive topics such as hatred, violence, violent extremism and terrorism online. The tertiary level can build on these approaches, providing further opportunities for young people to become involved in proactive activism and taking responsibility for enacting positive changes in their online communities.

Informal Education

Alongside the formal education sector, informal educational spaces and approaches can also contribute to preventing and countering the effects of violent extremism and terrorism online at the local community level. Informal education can help to reinforce the knowledge, skills, attitudes and behaviors young people learn in formal educational settings. Informal educational settings can often accommodate a wider and more flexible range of educational approaches, providing a vital avenue of support for young people that may struggle to learn in more formal settings, and more time and space for activities outside the core academic curriculum.

Involving Youth

Youth should not be viewed purely as vulnerable to violent extremism and terrorism online, but also as key stakeholders in designing and delivering effective educational and communications-based responses. Many young people demonstrate a desire to promote justice, effect positive change in the world and contribute to their communities and societies, and this energy, creativity and enthusiasm can be channeled in a positive manner to challenge the corrosive online narratives of violent extremist and terrorist groups. Young people are often aware of the conditions and drivers that lead their peers to radicalization and recruitment, and are also highly effective at communicating and influencing their peers and younger age groups.

Governments and educational institutions should therefore encourage young people to become active partners and youth mentors in efforts to prevent and counter violent extremism and terrorism online wherever possible. This could include facilitating interactions between youth and positive role models, or developing approaches where young people can raise awareness of online safety, violent extremism, or terrorism among their parents, families and communities.

Involving Parents & Adults

Comprehensive educational approaches should also involve engaging with parents, families and other adults to raise awareness of, and build resilience to, the dangers of violent extremism and terrorism online. Governments, educational institutions, civil society organizations and the private sector can collaborate to provide resources and training opportunities on safeguarding, online safety, preventing and countering violent extremism and terrorism online, and recognizing potential early warning signs of online radicalization.

Schools, as trusted institutions with existing relationships that are embedded in local communities, can act as the venue for such initiatives, which can also be incorporated within any existing programs designed to engage parents and families. When provided the necessary resources and training, parents and families can reinforce young peoples' learning from formal or informal educational settings at home.

Involving the Private Sector

Governments can encourage the private sector to establish Corporate Social Responsibility (CSR) programs to support educational responses. This may include providing expertise, resources and training to youth and other stakeholders, or supporting the dissemination of key online safety messages and promoting the availability of programs in this area. This could involve the creation of new programs, or the adaptation of existing online safety initiatives to include P/CVE relevant content and concepts. The private sector can also play a key role in reaching older audiences online, promoting critical thinking, civil discourse and online safety to audiences of all ages.

B: The Spectrum of Educational Responses

Violent extremists and terrorists attempts to radicalize and recruit online thrive where critical thinking, digital literacy and awareness of the dynamics of the online space are lacking. A wide range of educational approaches can contribute to building resilience and reducing vulnerability to violent extremism and terrorism online, while also developing the vital knowledge, skills, attitudes and behaviors of young people. While educational approaches may differ in terms of their primary goals, they often share many of the same objectives or learning outcomes, despite the fact that they are also highly context-specific.

Regardless of the type of approach employed, these forms of education are typically most successful when they employ an active and experiential style of learning rather than more traditional teaching methods and pedagogies. This may include simulations, games, group

exercises and practical activities, and combine a combination of different types of media to capture and retain the attention of young people.

For the purposes of this toolkit, educational responses are broadly divided into 'explicit' and 'implicit' approaches to preventing and countering violent extremism and terrorism online:

- **Explicit (or P/CVE specific) approaches** directly address the topics of violent extremism and terrorism, and are typically more suitable for secondary age youth (or above).
- **Implicit (or P/CVE relevant) approaches** indirectly address the underlying factors that can contribute to building resilience and reducing vulnerability to violent extremist or terrorist messaging and other online threats in young people. These approaches can be appropriately designed for youth of any age and all levels of education.

The table below (*Fig. 2*) outlines the range of educational responses, including the appropriate educational level, goals, objectives and learning outcomes for each type of response.

Figure 2: Explicit and Implicit Approaches to Educational Responses

	Explicit (P/CVE Specific)	**Implicit (P/CVE Relevant)**	
Type of Response	P/CVE education	Digital literacy & citizenship education	Online safety education
Educational Level	Secondary, tertiary	Primary, secondary, tertiary	Primary, secondary
Goals	Build resilience to violent extremism and terrorism	Encourage positive use the internet and social media and build resilience to violent extremism and terrorism online and other online threats	Encourage safe and effective use the Internet and social media
Objectives & Learning Outcomes (indicative)	Understanding and awareness dangers of violent extremism and terrorism online Critical thinking skills Media literacy & online propaganda Grooming & recruitment tactics Considering alternative viewpoints	Critical thinking skills Media & 'image' literacy* Internet architecture and online communication (e.g. echo chambers & filter bubbles) Collective responsibility & safeguarding peers online Considering alternative viewpoints Digital citizenship & activism	Online privacy and reputation ("digital hygiene") Managing online information and security Online manipulation Online relationships and bullying Self-image and wellbeing online

* "Image literacy should teach students about the power of images, and how images are emotion-based and not proposition-based. It should be stressed that images cannot prove something in the way words can prove a proposition. The influence of different typefaces, fonts, colours and visual styles, as well as the effect of accompanying music should be discussed so students understand their effects. Schools should teach students the standards different types of media have for the use of altered images so that they can better judge the validity of images" R. Hornik, *A strategy to counter propaganda in the digital era*, Yearbook of the Institute of East-Central Europe, 2016, Volume 14, Number 2, pp. 61–74.

As with other forms of communications-based responses, Governments should ensure that educational approaches are designed based on existing empirical evidence to ensure new curriculum developments or programs to prevent or counter the impacts of violent extremism and terrorism online are effective. This may include baseline research, such as needs-assessments, perceptions' studies, analyses of existing education literature and online statistics, and the development and assessment of pilot programs.

Examples of explicit P/CVE education programs include Bounce, an educational program, funded by the European Commission and coordinated by the Belgian FPS Home Affairs, which builds young people's resilience to violent extremism. [153] A second example is Extreme Dialogue, a project funded by Public Safety Canada through the Kanishka Project[154], and then through the European Union's ISEC program.[155] Extreme Dialogue is an interactive education resource for parents, teachers and youth workers that centers on compelling films, including of former violent extremists and survivors of violent extremism in the UK, Canada, Germany and Hungary.[156]

Finally, in 2016 Russia's *National Center for Information to Counter Terrorism and Extremism in the Educational Environment and the Internet* launched an online resource that provides information on planned activities around the country to promote active citizenship among children and youth. Additionally, the Zero Discrimination program was launched in the lead up to the 2018 FIFA World Cup to improve knowledge and reduce the risk of extremist and discriminatory actions by young people aged 14–21 by strengthening humanist values using examples from sports and football. The program adopts an interactive approach, centered around short topical videos and a series of participatory activities, including group assignments and discussions.

C: Implementing Educational Responses

Due to the fact that educational approaches to preventing and countering violent extremism and terrorism online are relatively new, Governments will likely face challenges in effectively scaling and mainstreaming them throughout the education system. Partnerships with education and youth institutions, civil society and the private sector are crucial in building the scale and reach of such approaches. Governments should carefully consider the existing skills, needs, and requirements of youth-focused practitioners when determining their approach.

Safe Spaces

In order for educational approaches to be effective, schools and other educational institutions and settings should be maintained as "safe spaces" where ideas can be freely expressed, discussed and debated in a non-judgmental fashion, free from discrimination, harassment, or threats of or actual emotional or physical harm. This type of approach should be established

153 See https://www.bounce-resilience-tools.eu/.
154 See https://www.publicsafety.gc.ca/cnt/ntnl-scrt/cntr-trrrsm/r-nd-flght-182/knshk/index-en.aspx.
155 See https://ec.europa.eu/home-affairs/financing/fundings/security-and-safeguarding-liberties/prevention-of-and-fight-against-crime_en.
156 See https://extremedialogue.org/.

and normalized as an institutional ethos to allow a range of perspectives to be explored, and any grievances to be addressed in an open and safe manner.

Educational and youth-focused institutions should consider providing training to their teachers or staff around how to engage in discussions with young people on sensitive topics safely and effectively, and how to avoid further alienating or increasing the vulnerability of any individual. Similarly, in contexts where youth may have been exposed to trauma or violence (e.g. refugee or conflict/post-conflict populations), teachers or staff should be mindful of the impact that this may have had, and account for this in their teaching methods and approaches.

Building on Existing Skills & Resources

In terms of addressing sensitive topics, P/CVE relevant training can be related and compared to existing online social issues or threats that are already familiar to teachers and youth-focused practitioners. Depending on the context, this could include relating the issues of violent extremism and terrorism online to other issues already covered by practitioners, such as gang violence and crime, drugs and alcohol, trauma, grooming and cyber-bullying. Such an approach helps to assure practitioners that preventing and countering violent extremism and terrorism online may require additional knowledge and understanding, but that they already possess many of the necessary skills.

Training and resources for the wider educational sector should also be considered to secure buy-in for P/CVE approaches throughout the system. This could include a wide range of stakeholders, including school leaders, management or administrators, state or local officials (e.g. from Ministries of Education, Culture, Youth, Sport, Religion), education training providers, school assessors or regulators, academics and researchers, and professional bodies. Training for these groups may include a more basic overview of the online threat, the goals and objectives of P/CVE specific or relevant approaches, key terminology, and the roles of different stakeholders in delivering a joined up whole-of-society educational response to violent extremism and terrorism online.

Curriculum Integration

Governments (either national, regional or local, depending on the education system) are typically well-placed to integrate changes to curricula to enable comprehensive educational responses to violent extremism and terrorism online to reach all young people through formal education. Depending on the extent to which related areas are present already, Governments can augment and expand subjects that emphasize or relate to citizenship education, shared values or human rights with P/CVE relevant content and learning outcomes. Integrating P/CVE content into existing subject areas can help to reduce the burden on teachers by relating a sensitive topic to areas they are comfortable delivering, and also avoid overcrowding the syllabus and adding to the pressures on teachers' time.

Monitoring & Evaluation

Monitoring and evaluation is vital to demonstrate the results and impact of programs, and can help to secure buy-in from key stakeholders (*see Chapter 4 B: Monitoring and Evaluation*

Communications-Based Responses).Where gaps in the evidence-base exist, Governments should conduct or support further multi-disciplinary research to identify best practices and improve responses on an ongoing basis, as well as iteratively adapting responses as the digital and threat environment evolves. Given the relative lack of evaluated programs in this area, best practices may be derived from other areas or fields, including theories of learning or pedagogy. Governments should also ensure that there are the correct incentives to encourage education providers to evaluate and critically assess their programs, and that the outcomes are shared across the sector.

Case Study: Preventing and Countering Extremism and Radicalisation – National Action Plan (Denmark)[157]

The Danish National Action Plan outlines a comprehensive approach to preventing and countering violent extremism and radicalization that brings together national and local authorities, various agencies, the education sector and civil society, with a particular focus on children and young people. Preventative approaches primarily aim to promote and safeguard the welfare and development of children by encouraging active citizenship, building democratic, social, critical thinking and employment skills, discouraging "risk behavior" and enhancing the resilience of youth to extremist messages.

This approach incorporates stakeholders from the national level (Ministry for Children, Education and Gender Equality, and the National Agency for Education and Quality), the local level (municipalities), local agencies (police districts, social services and "Info-houses" that provide expertise on extremism and radicalization), and the education sector (day-care facilities, primary and upper secondary schools, and youth and adult education programs). At the local level, stakeholders are brought together through SSP crime-prevention partnerships (Schools, Social services and Police). National government supports these efforts by conducting research, sharing expertise and knowledge, providing counselling and training, and developing and independently evaluating specific approaches and initiatives in order to share best practices.

Formal Education: Curriculum

The National Action Plan places specific focus on enhancing young peoples' critical thinking skills and understanding of citizenship through the national curriculum (Danish Folkeskole) objectives for primary and lower secondary schools. This includes the inclusion of human rights education in Social Studies (a mandatory subject covering health, relationships and family education), and an enhanced focus on digital and source literacy in Danish and History lessons. These topics are emphasized through an annual "theme week" that promotes the importance of

157 Udlændinge- og Integrationsministeriet, *Preventing and countering extremism and radicalisation National action plan*, last updated 15 March 2017.

democracy, community and citizenship throughout the education system, delivered by the Ministry for Children, Education and Gender Equality.

Formal Education: Training & Resources

To facilitate the additional emphasis placed on these topics in the curriculum, all levels of the education sector, and related stakeholders (such as municipalities), are provided with pedagogical and professional training and a range of resources to ensure effective implementation. This includes the provision of "learning consultants" through The National Agency for Education and Quality, who run series of events across Denmark covering best practices in the teaching of democracy and citizenship. A pilot project on the prevention of hate crime was also implemented in selected schools to design and test additional resources for tackling bullying, division, prejudice and stereotypes among young people. The project offered training in pedagogy and facilitating dialogue around sensitive topics to teachers and school leaders.

The National Agency for Education and Quality designed and circulates materials through a national learning portal (www.emu.dk) that offer teachers, school leaders and other practitioners with tangible resources to encourage the inclusion of approaches to prevent marginalization and radicalization and build resilience to extremist and terrorist messaging online. This includes resources on critical thinking, propaganda and manipulation, and online safety and digital literacy primary and secondary schools and extra-curricular clubs. Finally, approaches and resources for schools to encourage the involvement of parents are also available.

Informal Education, Youth & Civil Society

A variety of approaches are also included in the National Action Plan for informal educational settings and youth-focused civil society and religious organizations, as well as the provision of activities for young people. Training and resources are also provided for municipalities in how to effectively collaborate with local civil society and youth organizations and co-develop constructive opportunities to engage young people.

The Danish Agency for International Recruitment and Integration (SIRI) has established a national peer-to-peer dialogue initiative for young people (aged 18 to 35) to spur debates on important topics, encourage independence, and develop a sense of ownership of communities and belonging in society. The initiative covers a wide range of issues including; "identity, family relations, opportunities for self-expression, social control, honor-related conflicts, social participation, freedom and responsibility, rights and obligations, pro- and anti-social groups, discrimination and non-discrimination, images of friends and enemies, intolerance, [and] extremism." A partnership has also been developed between national government and a variety of

educational institutions to mobilize young people to push back against radicalization online through positive or alternative narratives by providing training in digital communications.

Alongside such initiatives, SIRI also offers capacity-building training for local civil society and youth organizations and practitioners to enhance their ability to deliver programs to prevent and counter extremism and radicalization, encourage positive participation in local communities and activities, and engage vulnerable or at-risk groups. To complement this training, educational resources covering critical thinking and digital literacy have also been developed with the Media Council for Children and Young People specifically for informal educational settings.

Further References

Chapter One: Development and Adoption of Content-Related Legislation and Policies

Australian Government, Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act, April 2019.

Government of the United Kingdom, Online Harms White Paper, April 2019.

Government of France, Creating a French framework to make social media platforms more accountable: Acting in France with a European vision, May 2019.

Chapter Two: Development of Transparency and Accountability Mechanisms

Committee of Experts on Internet Intermediaries, Study on the human rights dimensions of automated data processing techniques and possible regulatory implications, Council of Europe, 2017.

Conway, Maura and, Moign Khawaja, Suraj Lakhani, Jeremy Reffin, Andrew Robertson & David Weir, Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts, Studies in Conflict & Terrorism, October 2018.

Jourová, Věra, Code of Conduct on countering illegal hate speech online: Fourth evaluation confirms self-regulation works, European Commission, February 2019.

Pearson, Elizabeth, Online as the New Frontline: Affect, Gender, and ISIS-Take-Down on Social Media, Studies in Conflict and Terrorism, July 2017,

Chapter Three: Providing content-based responses through multi-stakeholder collaboration

Huang, Medea and Faris Natour, Legitimate and Meaningful: Stakeholder Engagement in Human Rights Due Diligence: Challenges and Solutions for ICT Companies, Business for Social Responsibility, September 2014.

Keen, Florence, Public–Private Collaboration to Counter the Use of the Internet for Terrorist Purposes: What Can be Learnt from Efforts on Terrorist Financing?' Royal United Services Institute for Defense and Security Studies, February 2019.

UNCTED, More Support Needed For Smaller Technology Platforms To Counter Terrorist Content, November 2018.

Tech Against Terrorism: Analysis, ISIS use of smaller platforms and the DWeb to share terrorist content, 2019.

Chapter Four: Development, Adoption and Evaluation of Policies

Bhulai, Rafia, Allison Peters & Christian Nemr, From Policy to Action: Advancing an Integrated Approach to Women and Countering Violent Extremism, Global Center on Cooperative Security and Inclusive Security, June 2016.

Cox, Kate, William Marcellino, Jacopo Bellasio, Antonia Ward, Katerina Galai, Sofia Meranto & Giacomo Persi Paoli, Social media in Africa A double-edged sword for security and development, UNDP & RAND Europe.

Directorate General for Internal Policies, Policy Department C: Citizen's Rights and Constitutional Affairs, The European Union's Policies on Counter-Terrorism Relevance, Coherence and Effectiveness, January 2017.

Feve, Sebastian and Mohammed Elshimi, Planning for Prevention: A Framework to Develop and Evaluate National Action Plans to Prevent and Counter Violent Extremism, Global Center on Cooperative Security, June 2018.

G7, G7 Action Plan on Counter Terrorism and Violent Extremism, October 2016.

GCTF, Ankara Memorandum on Good Practices for a Multi-Sectoral Approach to Countering Violent Extremism, 2013.

GCTF, Good Practices on Women and Countering Violent Extremism, 2014.

Hedayah, Guidelines and Good Practices for Developing National CVE Strategies.

Hedayah, Guidelines and Good Practices: Developing National P/CVE Strategies and Action Plans, September 2016.

Hedayah, The World of Communications Is the New Frontline in The Battle Against Violent Extremism.

Herrington, Rebecca, Emerging Practices in Design, Monitoring, and Evaluation for Education for Peacebuilding Programming, Search for Common Ground, October 2015.

IMPACT Europe, Innovative Methods and Procedures to Assess Counter-violent-radicalisation Techniques in Europe: Toolkit Manual.

Khalil, James & Martine Zeuthen, Countering Violent Extremism and Risk Reduction: A Guide to Programme Design and Evaluation, Royal United Services Institute, June 2016.

RAN Centre of Excellence, Developing a local prevent framework and guiding principles - Part 2, November 2018.

RAN Centre of Excellence, Monitoring & Evaluating counter- and alternative narrative campaigns, February 2019.

Russell, Olivia, Meet Me At The Maskani: A Mapping of Influencers, Networks, and Communications Channels in Kenya and Tanzania, Search For Common Ground, June 2017.

Tuck, Henry & Louis Reynolds, The Counter-Narrative Monitoring & Evaluation Handbook, 2017.

UNDP & International Alert, Improving the impact of preventing violent extremism programming: a toolkit for design, monitoring and evaluation, 2018.

United Nations General Assembly, Plan of Action to Prevent Violent Extremism Report of the Secretary-General, December 2015.

United Nations Office of Counter-Terrorism, Developing National and Regional Action Plans to Prevent Violent Extremism.

United Nations Security Council, Letter dated 26 April 2017 from the Chair of the Security Council Committee established pursuant to resolution 1373 (2001) concerning counter-terrorism addressed to the President of the Security Council, April 2017.

United Nations Security Council, Resolution 2354, May 2017.

Chapter Five: Collaboration with ICT Industry and Engagement with CSOs

Bhulai, Rafia, Going Local: Supporting Community-Based Initiatives to Prevent and Counter Violent Extremism in South and Central Asia, December 2017.

Committee on Legal Affairs and Human Rights Council of Europe Parliamentary Assembly, Counter-Narratives to Terrorism, March 2018.

Department of Homeland Security, Strategic Implementation Plan for Empowering Local Partners to Prevent Violent Extremism in the United States, October 2016.

Elsayed Lilah, Talal Faris & Sara Zeiger, Undermining Violent Extremist Narratives in the Middle East and North Africa: a how-to guide, Hedayah, December 2017.

GCERF, A Youth Perspective on Preventing Violent Extremism.

GCTF, Ankara Memorandum on Good Practices for a Multi-Sectoral Approach to Countering Violent Extremism, 2013.

Hemmingsen, Ann-Sophie & Karin Ingrid Castro, The Trouble with Counter-Narratives, Danish Institute for International Studies, 2017.

ICCT & Hedayah, Developing Effective Counter-Narrative Frameworks for Countering Violent Extremism, September 2014.

OSCE, The Role of Civil Society in Preventing and Countering Violent Extremism and Radicalisation that Lead to Terrorism: A Focus on South-Eastern Europe, August 2018.

RAN Centre of Excellence, A Nimble (NMBL) Approach to Youth Engagement in P/CVE.

RAN Centre of Excellence, Developing counter- and alternative narratives together with local communities, October 2018.

RAN Centre of Excellence, Guidelines For Young Activists: How To Set Up A P/CVE Initiative - Part 1: How to develop your own PVE initiative, March 2019.

RAN Centre of Excellence, Guidelines For Young Activists: How To Set Up A P/CVE Initiative – Part 2: How to develop a project plan for your P/CVE initiative, March 2019.

Reed, Alastair, Haroro J. Ingram & Joe Whittaker, Countering Terrorist Narratives, European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs, November 2017.

Zeiger, Sara, Counter-Narratives For Countering Violent Extremism (CVE) In South East Asia, Hedayah, May 2016.

Zeiger, Sara, Undermining Violent Extremist Narratives in East Africa: A How-To Guide, Hedayah, August 2018.

Chapter Six: Empowering Youth and Building Resilience through P/CVE, Online Safety, and Digital Citizenship

Centre on Global Counterterrorism Cooperation, The Role of Education in Countering Violent Extremism, December 2013.

GCTF & Hedayah, Abu Dhabi Memorandum on Good Practices for Education and Countering Violent Extremism.

MediaSmarts, Use, Understand & Create: A Digital Literacy Framework for Canadian Schools, 2019.

National Society for the Prevention of Cruelty to Children (NSPCC), https://learning.nspcc.org.uk/.

RAN Centre of Excellence, Handbook on CVE/PVE training programmes: Guidance for trainers and policy makers, December 2017.

RAN Centre of Excellence, Education and radicalisation prevention: Different ways governments can support schools and teachers in preventing/countering violent extremism, May 2019.

RAN Centre of Excellence, Transforming schools into labs for democracy: A companion to preventing violent radicalisation through education, October 2018.

UNESCO, A Teacher's Guide on the Prevention of Violent Extremism, 2016.

UNESCO, Global Media and Information Literacy Assessment Framework: Country Readiness and Competencies, 2013.

UNESCO, Preventing violent extremism through education: a guide for policy-makers, 2017.

UNESCO, Youth and Violent Extremism on Social Media, 2017.

UNESCO & Mahatma Gandhi Institute of Education for Peace and Sustainable Development, Youth Led Guide on Prevention of Violent Extremism Through Education, 2017.

United Network of Young Peacebuilders & Search for Common Ground, Translating Youth, Peace & Security Policy into Practice: Guide to kick-starting UNSCR 2250 Locally and Nationally, November 2016.

UK Department for Education, Educate Against Hate.

UK Home Office, E-Learning Training on Prevent.

UK Department for Education, Thomas Chisholm & Alice Coulter (Kantar Public), Safeguarding and Radicalisation Research Report, August 2017.

GCTF

GLOBAL COUNTERTERRORISM FORUM