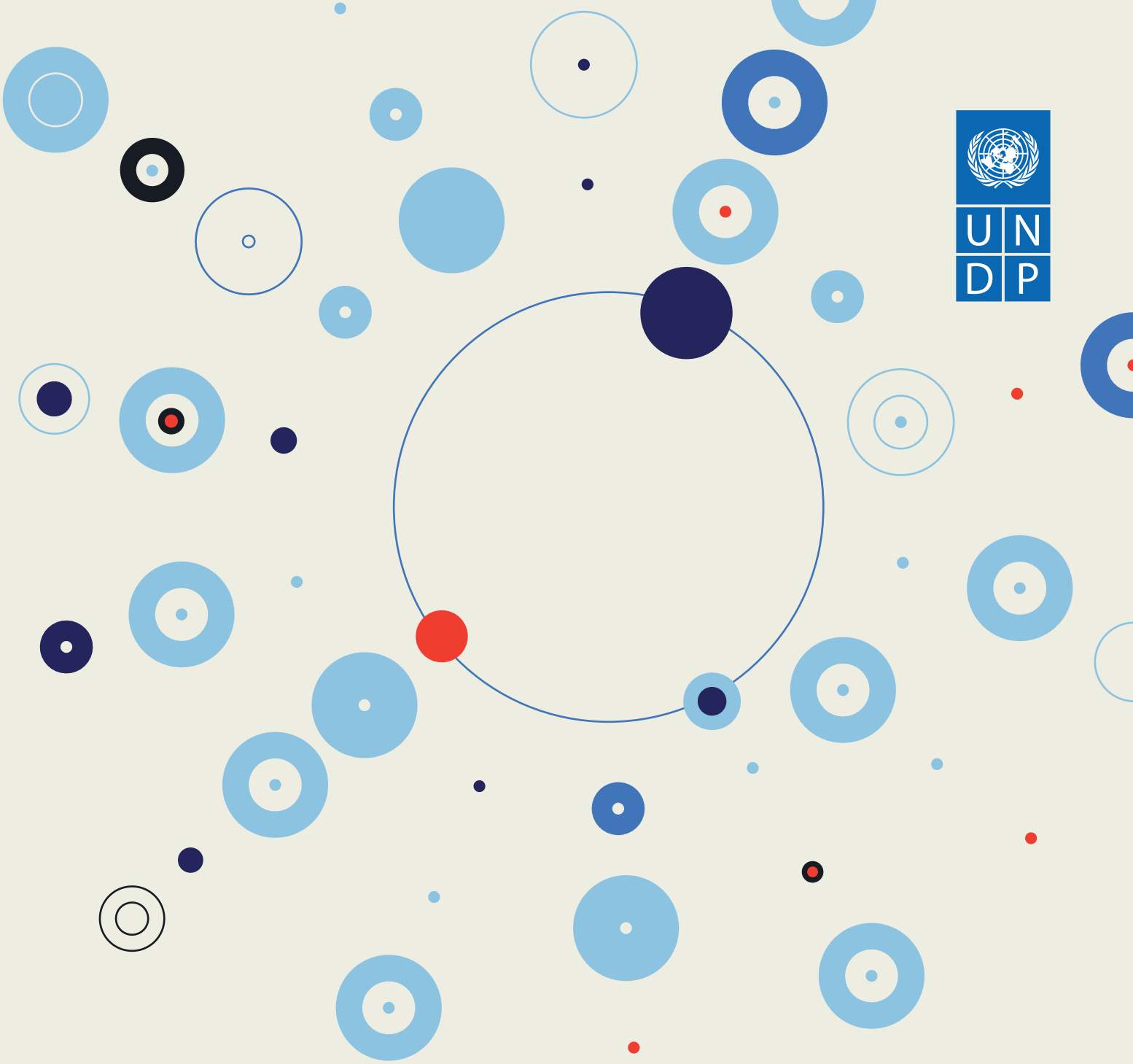


GUIDANCE NOTE

From Pilots to Practice:

Risk-informed utilization of online data for preventing violent extremism and addressing hate speech

One United Nations Plaza, New York, NY 10017, USA

The United Nations Development Programme (UNDP) is the leading United Nations organization fighting to end the injustice of poverty, inequality, and climate change. Working with our broad network of experts and partners in 170 countries, we help nations to build integrated, lasting solutions for people and planet. Learn more at undp.org or follow at @UNDP.

Acknowledgements

This guidance note was prepared in a process led by the Conflict Prevention, Peacebuilding and Responsive Institutions (CPPRI)/Prevention of Violent Extremism (PVE) Team at UNDP's Crisis Bureau. Under the editorial direction of Nika Saeedi and the guidance of Samuel Rizk (PhD), the development of the guidance note was supported by the lead author Angharad Devereux and contributing author Heesu Chung, with additions from Lucy Turner and Gitte Nordentoft. The team is grateful to those who contributed and are acknowledged in the sister publication, the policy brief "From Pilots towards Policies: Utilizing online data for preventing violent extremism". It is also grateful to the UNDP Global Policy Network (GPN) of PVE practitioners, Regional Hubs and Country Offices, UNDP Chief Digital Office, UNDP Crisis Bureau's Crisis and Fragility Policy and Engagement team, UNDP CPPRI / Conflict Prevention & Peacebuilding Team and UNDP Oslo Governance Centre. Particular thanks go to UNDP Bangladesh, Pakistan, Kyrgyzstan, Sudan, Iraq and Thailand offices for sharing their experiences and insights. Gregory Connor, Rob Stoleman and Mitra Modaressi of UNDP and Janeen Fernando of the UN Resident Coordinator's Office (RCO) in Sri Lanka also provided valuable feedback to the document. The Team would like to thank the Cyber Threats Research Centre of Swansea University, particularly Joe Whittaker (PhD), and Connor Rees, for their collaboration in this process and peer review of the document. The authors would also like to thank Erin Saltman (PhD), of The Global Internet Forum to Counter Terrorism (GIFCT) and Anne Craanen of Tech Against Terrorism for their contributions of expertise to the document.

Copy editor: Barbara Hall

Design and layout: Benussi & the Fish

For queries on UNDP's work in PVE, please contact Nika Saeedi: **nika.saeedi@undp.org**.

Programming Guidance

Table of Contents

Executive summary

An increase in the use of the internet across the globe has created opportunities in gaining contextual insights into potential drivers of violent extremism (VE) that can be utilized to better target programmes that aim to prevent the appeal of VE. UNDP preventing violent extremism (PVE) practitioners across seven country offices have been undertaking pilot projects that collect, analyse and apply online data to programming for PVE and addressing hate speech. They have found positive potential in creating a richer evidence base than when utilizing offline data collection methods alone to inform programming. However, the monitoring of online data, particularly content from social media platforms, is a relatively new opportunity in data collection for PVE practitioners and therefore involves a range of practical, ethical and rights-based challenges. These challenges include the time and expertise necessary to undertake a project that collects and analyses online data, the complex policy and stakeholder landscape around the availability of data, and the ability to ensure consistent application of human rights to this relatively new sphere online. This guidance note aims to set a framework for PVE practitioners who may be planning to utilize online data to inform PVE programmes, helping guide them through considerations to make when forming their own project methodologies and risk assessments in order to fully unlock the potential of risk-informed use of online data for PVE. In this context, this guidance note promotes risk management as an inherently enabling process, which is encouraged in order to unlock the greatest potential of enhancing the evidence base of PVE programming with online data.

Checklist of key considerations when using online data for preventing violent extremism

OBJECTIVE, METHODOLOGY DESIGN AND RISK ASSESSMENT OF ONLINE DATA

CONSIDERATION 1

- Determine whether it is useful or not to monitor online data for preventing violent extremism (PVE) programming and policy, as opposed to or in addition to more traditional data-gathering methods to reach desired objectives.
- Design realistic objectives related to using online data to enhance organizational understanding of the uses of the internet by violent extremist groups for informing PVE programmes and policy.
- Develop relevant and effective methodologies that are suitable for monitoring and analysing available data and that help meet project objectives with regard to addressing VE-related use of the internet.

CONSIDERATION 2

- Enhance organizational coordination and capacity in data usage, including objective and methodology setting and subsequent risk assessment, data collection, analysis, application, and monitoring and evaluation.

CONSIDERATION 3

- Develop an understanding of potential risks from experience and available resources in order to inform risk assessments focused on the human rights implications of projects that utilize data for PVE.

CONSIDERATION 4

- Undertake due diligence for potential partners, framed around the potential benefits and risks (including risk mitigation options) that may emerge when working with the specific type of partner (e.g. private organization, social media company, non-governmental organization, or civil society organization).
- Look for opportunities to engage local-level actors where possible.

COLLECTION, ANALYSIS, AND APPLICATION OF ONLINE DATA

CONSIDERATION 5

- Develop understanding of data made available to the organization to determine the most effective ways to use this for the project's objectives.

CONSIDERATION 6

- Complement the application of online activities with offline activities aimed at tackling offline networks and relationships, or providing positive alternative spaces for social and activist engagement.

CONSIDERATION 7

- Develop an understanding of the free or commercial tools available for data usage as well as an understanding of where and how the development of new tools may be more advantageous than using existing ones. Develop an understanding of where AI can and cannot be of use to a project and how it can be utilized ethically and effectively.

CONSIDERATION 8

- Ensure that a communication plan is in place to effectively communicate relevant findings with appropriate stakeholders in order to ensure the maximum impact of learning acquired.

MONITORING, EVALUATION, AND LEARNING

CONSIDERATION 9

- Adapt and update monitoring and evaluation (M&E) frameworks to incorporate online data from new sources, including online platforms, as part of the evidence base to monitor progress and evaluate the impacts of PVE projects.
- Use M&E processes to ensure that objectives are met and methodologies are updated where deemed ineffective through iterative and adaptive programming.

CONSIDERATION 10

- Ensure that learning feeds into future programming and policy, including that of partners and stakeholders, in order to enhance the practice of risk-informed data utilization for PVE across a 'whole-of-society' approach.

Acronyms and abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
CRD	Crisis Risk Dashboard
CSO	Civil Society Organization
DPPA	Department of Political and Peacebuilding Affairs
GDPR	General Data Protection Regulation
GIFCT	Global Internet Forum to Counter Terrorism
M&E	Monitoring and Evaluation
NAP	National Action Plan
P/CVE	Preventing and Countering Violent Extremism
PII	Personally Identifiable Information
PVE	Preventing Violent Extremism
SDG	Sustainable Development Goal
TCAP	Terrorist Content Analytics Platform
UDHR	Universal Declaration of Human Rights
UNDP	United Nations Development Programme
VE	Violent Extremism

@UNDP Bangladesh

Introduction

The prevalence of the internet in everyday lives has expanded, albeit unevenly, across the globe, with a particularly notable uptake in use of social media. While the internet creates many positive opportunities for users, this increased uptake also creates the possibility that users may come into contact and interact with harmful material online, including material that perpetuates violent extremist (VE) narratives.

Therefore, the use of online data, and potentially artificial intelligence (AI) tools to scale up data collection and analysis of VE trends online has received increasing attention from policymakers and practitioners. This is in part due to the opportunities in gaining live insights to better target programmes that aim to prevent the appeal of VE. UNDP practitioners of preventing violent extremism (PVE) have found that online data offers the potential to better understand the drivers of radicalization and hate speech that can lead to VE in certain contexts. Clear solutions, however, are made more challenging by the complex policy and rights landscape surrounding the monitoring of online data for development-based purposes.

Considering these trends and developments, UNDP has been capturing learning from pilot projects that have been monitoring online data to inform the planning and implementation of development-based policy and programmes for PVE. This has also been undertaken due to the limited nature of publicly available information on utilizing this opportunity in a risk-informed manner due to the relatively limited global experience in the use of these technologies for these purposes.

This guidance note aims to introduce the learning acquired from UNDP PVE practitioners on utilizing online data for PVE and addressing hate speech for global practitioners to understand the opportunities and challenges in this area and therefore create stronger projects and programmes. The note, as a sister product to the policy brief, 'From Pilots towards Policies: Utilizing online data for preventing violent extremism and addressing hate speech', provides an overview of key processes, tools and resources to support the practitioner with practical guidance. Given their introductory nature, these communication products aim to lay the foundations for informed policy and guidance from practice and are by no means intended to be an exhaustive overview of the application of online data, including the use of online and more developed AI tools, to prevent VE.

Pilot projects in use of online data for preventing violent extremism

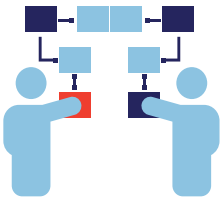
A pilot project here is an initial targeted implementation of using new technologies to collect, analyse and apply data, with learnings used to test the viability of a project idea, and therefore inform future programming across the organization using similar approaches. This could involve the application of a standard methodology recommended, or implemented alongside, outside parties but which is relatively new to the organization in comparison to more traditional and more heavily offline-based data collection methods.

The first section of this guidance note introduces the foundational aspects that should be incorporated into projects at the design stage in order to strengthen the impact and reduce the potential of harm. These aspects include ensuring that an organization develops the technical capacity for teams to form effective and realistic objective-centred methodologies, to develop an ensuing risk assessment and engage partners where beneficial. The second section highlights considerations to make specifically for the collection of data, including the type of data available, the need to supplement online methods with offline, and when tools can aid data collection. The third section outlines the latter stages of the data cycle, offering suggestions on how to heighten the impact of findings and how the data process as a whole should be monitored and evaluated and how it can strengthen further iterations of data use for PVE. Reference to any specific organization, tool or application in this report should not be considered an endorsement by UNDP or by the United Nations. The tools and applications mentioned in this report are included solely to demonstrate the potential application of online data for PVE.

Identifying and implementing opportunities and addressing challenges and risks: processes for using online data and AI applications for PVE

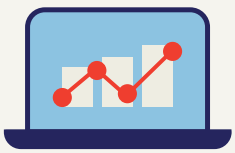
Informed by UNDP pilot programming, the data cycle outlined below identifies the main elements of PVE programming that utilizes online data.

Data Cycle



a) Objective, methodology design and risk assessment
To utilise online data, an ultimate objective must be decided on to guide the process. From this a methodology can be developed, informed by a risk assessment process.

b) Collection
Qualitative and/or quantitative data can then be collected from online sources using manual and/or automated tools.



e) Monitoring and Evaluation
M&E processes can help gain lessons learnt from each stage of the project, including a greater understanding of risks in order to inform future risk mitigation and strengthen methodologies going forward.

d) Application
The analysed data can then be communicated and applied to projects and programmes and/or inform policy.

c) Analysis
The data collected can then be validated, reflected upon and organised in a way that helps the team gain insights from the data in line with project objectives.

Data Cycle

a) Objective, methodology design and risk assessment

First, it must be determined whether online data can be useful for the project context and objective. Then, to begin the process of data collection, analysis and application, a methodology should be designed, risks ascertained, and processes to manage the data developed. Illustrative examples of objectives of data collection include to:

- Analyse the actors and trending narratives of hate speech in a local or regional context.
- Understand how violent extremism (VE) narratives build on current affairs.
- Understand the audiences that are most susceptible to VE narratives and how they are targeted.
- Understand the potential and active driving factors of VE.
- Enhance early warning systems.

A methodology for the project must be designed to meet the objectives with expertise necessary. Risk assessment should inform a cost-benefit analysis of projects, deeming whether the potential gains are greater than any potential practical and ethical risks. Human rights considerations should always frame these processes and should never be a cost.

b) Collection

Qualitative and/or quantitative data can be collected and cleaned (e.g. case transformation, tokenization, unnecessary symbol removal, hashtag processing) from online sources, such as social media platforms, blogs or online news outlets manually by individuals monitoring data sources and by using data collection tools. This process may then be automated by individuals who train algorithms to collect a potentially larger quantity of data in a pre-targeted manner through AI tooling. Automatizing data collection entails:

- selection and filtering using a codebook or lexicon can be developed for humans and potentially machine-learning tools (found within AI tools) to track the most relevant data from the online sources chosen to be most relevant, and which data is available to gather from, based on a given context and objective(s).

However, there are challenges to new algorithm and AI tool development, including the need for highly technical expertise such as those offered by tech platforms, engineers and data scientists. Therefore, it may be more practical for PVE practitioners to first collect data with qualitative samples in order to build a more accurate understanding of specific target audiences. Data collection can then be scaled as required using various online social listening tools, mapping tools or tools that match similar audiences. The limitation of utilizing existing tools is that the languages and scripts to be monitored may not be offered by tooling. Due to higher privacy models of social media companies, a partnership with a platform may be deemed necessary after undertaking due diligence for assessing whether a potential partnership/partner is beneficial/ethical.

c) Analysis

In order for the data to be communicated clearly, data visualizations should be accompanied with meaningful narratives in order to humanize the statistics at hand. These information products can be built upon the needs of the user.

Four analysis methods that can be used, depending on what insights aimed for are:

- Quantitative discourse analysis – What are people talking most about?
- Netnography or narrative analysis – How are people talking?
- Sentiment analysis – What is the tone and emotion in a narrative?
- Network analysis – Who is talking about what with whom?

Data can be used to develop various visualizations and dashboards in order to have data organized in a way that is readily accessible and updated.

d) Application

The analysed data can then be interpreted for patterns and utilized for programming by applying findings to objectives such as to understand VE drivers and narratives, targeting at-risk audiences and informing alternative narratives and early warning systems, etc., for enhancing PVE programming. Data can also be utilized to build the evidence base for advocacy with tech companies for policy impact, which can be particularly impactful if collated in a sufficiently comparable manner. Application of online data can either be via online means (such as alternative narrative or digital literacy campaigns) or via the many offline PVE activities that may benefit from these data-informed approaches. The analysed data would then be visualized, disseminated and shared with the necessary stakeholders who can utilize them, for example, government counterparts, civil society organization (CSO) partners, programming staff and tech companies while taking into account safety, security and other risk considerations.

e) Monitoring and Evaluation

Monitoring and evaluation (M&E) frameworks and tooling can draw lessons learned, including a greater understanding of risks and possible impacts in order to inform future risk mitigation and strengthen methodologies going forward.

@UNDP Bangladesh

OBJECTIVE, METHODOLOGY DESIGN AND RISK ASSESSMENT OF ONLINE DATA

CONSIDERATION 1:

Determining the objective and designing an effective and relevant methodology

What types of information are created by VE groups and individuals online for PVE practitioners to monitor?

First, the suitability, applicability and objective of enhancing offline data collection methods with online data collection, analysis and application must be determined together with an appropriate methodology. In order to determine whether or not to use online data, practitioners must begin with an understanding of why online data is worth monitoring to inform PVE projects and programmes (for a discussion of the true potential of online data, see also *Consideration 5* on accessing relevant data). Then, the type of online data that will fill in gaps in understanding of issues related to PVE should be determined. This will contribute towards building an effective and relevant methodology, which includes the identification of the right data, tools and teams to reach a realistic objective.

Based on the defined objective, data source(s) should be identified and the methodology for online data collection and analysis, as further discussed in Section II and III of this paper, should be developed. The data available to PVE practitioners will likely be on the open web, which is allowed by platforms' Terms of Services, including social media data. PVE practitioners often utilize aggregate data and information on the general sentiments of an online community in order to understand audiences, narratives, drivers, etc., rather than specific data on an individual actor, which can lead to breach of privacy. To ensure safe, ethical and secure use of data, it is necessary to understand the purpose of the materials available on the open web, as opposed to trying to gain access to encrypted or private channels, file-sharing sites, or the Dark Web, for example.

The most relevant tactic of VE groups to PVE practitioners, which informs much of VE activity on the open web, is the development of community. While offline relationships are extremely relevant to understanding the pull of violent extremism, the online elements of radicalization, such as the development of online narratives through propaganda, and related to this, the development of online communities, can help facilitate the path of radicalization. Understanding how communities are formed online, including how offline factors are utilized, can help practitioners form relevant project objectives based on the kind of data that is available from this type of use of the internet.

Development of online communities

Violent extremist group identity is created by linking concepts of crisis or threat to the out-group and the solution to this to the in-group in order to strengthen identity on both sides.[1] This crisis perception encourages automatic, non-deliberative thinking, which in turn lends itself to encouraging individuals to believe and possibly act on views that sway far from normality.[2] Once an 'us vs. them' dynamic is developed, individual efforts to help the 'us' often lead to actions against the 'them',[3] and once individuals become 'fused' (i.e. extremely identified) with their group, they become more willing to fight or die for them.[4] Being motivated by love of a group can facilitate viewing the other side as motivated

1 Tajfel, H. (1979). Individuals and Groups in Social Psychology. *British Journal of Social and Clinical Psychology*, 18; Ingram, H. (2016). A "Linkage-Based" Approach to Combating Militant Islamist Propaganda: A Two-Tiered Framework for Practitioners. *Terrorism and Counter-Terrorism Studies*; Berger, J.M. (2017). Extremist Construction of Identity: How Escalating Demands for Legitimacy Shape and Define In-Group and Out-Group Dynamics. The International Centre for Counter-Terrorism – The Hague 8, no. 7.
2 Ingram, H. (2016). Deciphering the Siren Call of Militant Islamist Propaganda: Meaning, Credibility & Behavioural Change. *Terrorism and Counter-Terrorism Studies*.
3 Ellemers, N. (2012). The Group Self. *Science 336*.
4 Swann, William B., and others (2008). Identity Fusion: The Interplay of Personal and Social Identities in Extreme Group Behavior. *Journal of Personality and Social Psychology*, 96.

Objectives commonly identified: What do we mean by use of online data for PVE in practice?

Monitoring online data can help practitioners identify prominent VE narratives and/or hate speech trends (through netnography or narrative analysis), and how they are being interpreted online (what people are talking about most and how). This data can also help determine who are the prominent online actors in a given context, and the types of online engagements they are involved in. Determining the end use of data will inform decisions on what insights one aims to gain from the data collection and analysis process. For example, if data is being monitored to inform effective counter-narratives, tracking online data can help practitioners understand whom to target online, what type of engagement to target them with, and how to adapt to local contexts with appropriate cultural references.

The identified objective can then help decide what type of data needs to be collected. Data can be quantitative (e.g. what people are talking about most) or qualitative (e.g. what is the tone and emotion [sentiment analysis] of a narrative and who is talking about what with whom [network analysis]). Some examples of UNDP objectives for monitoring online data for PVE are to:

- **Understand how VE narratives build on current affairs in order to better target PVE programming:**
  For example, UNDP Pakistan aimed to understand the preponderance and resilience of hyper-masculine narratives in social and digital media in Pakistan. The objective of this project incorporated the analysis of their social impacts, informing possible policy and programming actions towards the safety, inclusion, equality and empowerment of women, girls and gender minorities in national development, in line with Sustainable Development Goal (SDG) 5.

- **Understand the audiences that are most susceptible to VE narratives (including how these audiences interpret VE material online, potentially in comparison to offline, and how they are targeted by VE groups):**
  For example, UNDP Bangladesh aimed to monitor a range of violent extremist narratives and associated online harms in cyberspace. In doing so, this work outlined what the most salient issues to online communities of Bangla-speaking sympathisers are and gained a better understanding of numerous contributing factors including whether economic inequality, development, or human rights concerns in Bangladesh or among the Bangla-speaking diaspora shape violent or exclusionary narratives online.

- **Enhance early warning systems and/or to understand the potential and active driving political, societal and economic factors of VE in a given context:**
  For example, UNDP's Regional Bureau for Arab States' Digital Lighthouse Initiative: 'Applying Big Data and AI in the context of Hate Speech across Social Media' leveraged knowledge from the social sciences on the antecedents, root causes and predictors of hate speech, and detected and predicted these precursors in instances of hate speech on Twitter.

by hate,[5] making it easier and more desirable to harm individuals of another perceived group, or 'out-group',[6] since less empathy and more pleasure is felt in their pain.[7] Therefore, the themes and topics of narratives used by VE groups are relevant for PVE practitioners to monitor because they can contribute to heightening in-group mentality. Understanding and tackling these topics and themes can then disrupt the strength of VE attempts to 'radicalize' individuals.

This dynamic is strengthened by the system of meaning that extremists create for members: new lenses by which to interpret the world around them in line with group norms.[8] VE groups will directly or indirectly refer to an 'us versus them' narrative through propaganda that can take the form of formal or informal text, image, or video, and will often use and manipulate historical or current events to legitimize an argument. Therefore, it is worth considering the different formats VE narratives may utilize online at methodology design stage, which may vary particularly with

5 Waytz, A., L. Young, and J. Ginges (2014). Motive Attribution Asymmetry for Love vs. Hate Drives Intractable Conflict . Proceedings of the National Academy of Sciences, 111.
6 Reicher, S., Alexander Haslam and Rakshi Rath (2008). Making a Virtue of Evil: A Five-Step Social Identity Model of the Development of Collective Hate. *Social and Personality Psychology Compass*, 2.
7 Cikara, M., and others (2014). Their Pain Gives Us Pleasure: How Intergroup Dynamics Shape Empathic Failures and Counter-Empathic Responses. *Journal of Experimental Social Psychology*, 55.
8 Ingram, H. (2016). Deciphering the Siren Call of Militant Islamist Propaganda: Meaning, Credibility & Behavioural Change. *Terrorism and Counter-Terrorism Studies*.

platform, as well as the different relevant aspects of the data to organize such as content, topic, actor and sentiment. Local partners can help choose those most relevant platforms to a given context, objective (e.g. certain topics may be more likely to be discussed on certain platforms) or target audience (demographics of platforms may vary). DataReportal, for example, compiles free reports from third-party, trusted data sources and hosts profiles for most countries with statistics on internet penetration and social network use.[9]

The development of collective hate – the feeling of hatred against out-groups that could yield a celebration of inhumanity against those out-groups – has been pinpointed to five stages.[10] The internet offers VE groups the chance to reach more individuals and use multimedia propaganda to heighten the potential of users moving further along the stages of collective hate. PVE practitioners can utilize online data to understand the narratives that drive these stages in order to prevent their appeal. These stages can be used to categorize VE narratives online when analysing collected data in order to track the potential strategy and success (through for example views, such as through clicks and reshares, and sentiment, such as through comments and likes/dislikes) of VE groups to further move individuals through the development of collective hate.

1. **Identification**: This refers to the creation of a cohesive in-group that shares a common social identity that prescribes a set of norms and beliefs associated with their social categorization.
2. **Exclusion**: The creation of an inclusive in-group identity also involves creating boundaries around group membership.
3. **Threat**: Once targets are excluded from the in-group category, this distinction between 'us' and 'them' is made salient. The key factor that creates out-group hostility is not categorization per se, but rather the perception of the out-group as posing a threat to the in-group.
4. **Virtue**: The most hostile intergroup contexts often involve groups who consistently extol their in-group as uniquely virtuous and good, while positioning out-groups as lacking this same moral stance and virtue, and in fact threatening the values and norms of the in-group.
5. **Celebration**: The final stage in the experience of collective hate identified by Reicher et al. occurs when hostility and expressed hatred (both symbolic and physical, be it hate speech, hate crimes, or outright genocide) against threatening out-groups are celebrated, glorified and extolled. Hate becomes normalized, expected and almost prescribed within the in-group identity.

In-group terms, symbols and images are key to establishing a community, forming common value and norm systems beyond geographic limitations. A move to encrypted online communities can be particularly powerful as members willingly choose to participate and interact with material rather than simply receiving it on the open web.[11] Validation for ideas can flourish, and interpersonal relationships can form around online discussions, which create potential for recruitment.[12] Recruiters are known to study posting behaviour in order to select those who would be most likely to act for the organization.[13] In order to reduce the strength of these interactions through mitigating programmes and policy, it is necessary to understand the nature and driving forces behind them to reduce their use for VE groups.

The relevance of monitoring hate speech for preventing violent extremism

Understanding the drivers of hate speech in communities is crucial to developing interventions that can limit and repair the harm caused by hate speech. This understanding can also help identify those most susceptible to producing, and at risk from receiving, hate speech and enable their participation in dialogue on hate speech and appropriate counter measures. There is a growing body of evidence linking hate speech, especially online, to radicalization, which can lead to violent extremist acts in a small, but significant minority of those exposed to such negative content.*

Note:

* See, for example: Muller, K., and C. Schwartz (2020). *Fanning the Flames of Hate: Social Media and Hate Crimes* (2020) SSRN. Available here; Waldron, J. (2012). *The Harm in Hate Speech* (Harvard University Press. Boston, www.jstor.org/stable/j.ctt2jbrjd; and Government of Scotland (2018). Independent Review of Hate Crime Legislation in Scotland. Available here.

9 DataReportal, https://datareportal.com
10 Reicher, S., Alexander Haslam and Rakshi Rath (2008). Making a Virtue of Evil: A Five-Step Social Identity Model of the Development of Collective Hate. *Social and Personality Psychology Compass*, 2.
11 Bowman-Grieve, L. (2009). Exploring "Stormfront": A Virtual Community of the Radical Right. Studies in Conflict & Terrorism 32, p. 99.
12 Sageman M. (2011). *Understanding Terror Networks*. University of Pennsylvania Press, Inc.
13 Galloway B., and R. Scrivens (2018).The Hidden Face of Hate Groups Online: A Former's Perspective. *VoxPol*.

CONSIDERATION 2:
Building technical capacity

How can practitioners gain the type of understanding and skills necessary to undertake projects that utilize online data for PVE?

In-house coordination and capacity building

To undertake a project that utilizes online data for PVE, a multidisciplinary team is required together with considerable time and flexible resources available throughout the lifecycle of a project. Therefore, it is important to determine any gaps in capacity for data collection, including in the development of methodologies, lexicons and IT skills, and whether in-house, and/or external capacity building of partners is needed.

UNDP Data Strategy: A framework for organizational coordination for the effective and ethical use of data

Built on UNDP's Data Principles, the strategy:*

- outlines the specific benefits of harnessing data to UNDP;
- summarizes the approach of the strategy, including how it aligns with other organizational plans related to data;
- outlines how data are already being used across the organization;
- clearly states how effective and ethical use of data to be prioritized, including through M&E of the strategy and the development of a centralized governance structure to coordinate data usage including implementation of a dedicated data team and surrounding network and group to increase digital literacy and ensure organization-wide representation and use cases to implement and learn from.

In order to strengthen data collection and analysis, the strategy sets out plans for:

1. **Introductory guides** for primary data collection (including design, research ethics and best practices on data management and analysis with an emphasis on gender-sensitive data)
2. **A depository of UNDP research projects** including research design, questionnaires, training materials and analysis.
3. **A centralized capability, infrastructure and technology** to harness and analyse data from new data sources (digital footprints, geospatial, satellite data, etc.) safely and responsibly.
4. **Tools for data collection** with governance rules embedded, standardised indicators (e.g. gender, SDGs). Training, tutorials and example form templates to be adapted.
5. **Integrated data analysis support** for summary/advanced statistics and data visualizations.

Source: UNDP Data Strategy (2021).
Note: *UNDP. Data Principles: 8 Data Principles for UNDP. Available here.

To build UNDP's organization-wide capacity in this area and enact the Secretary-General's United Nations Data Strategy, a data strategy (see box below) has been developed to drive value-oriented governance. The UNDP Data Strategy seeks to establish a framework to produce norms, standards and policies that guide how data is collected, stored, protected and shared with UNDP against new Strategic Plan outcomes. These standards and policies will provide guidance to enable staff and partners to be forward thinking in their data-related initiatives, both internal and external facing, and to ultimately support member states and deliver better services.

The strategy has three major pillars – governance, people and technology:

- The governance pillar seeks to ensure that the processes are in place to manage data effectively and ethically, and that there are clear lines of escalation for any issues that arise.
- The people pillar seeks to foster a workforce that is future-ready and that can harness data in daily workflows. Enabling a data-literate UNDP is at the core of the strategy.
- The technology pillar starts to invest in a corporate data architecture, providing UNDP with cutting-edge self-service storage, distribution and analysis tools.

UNDP's Chief Digital Office has also developed the Digital X Scale Accelerator to scale up impact and geographic reach of successful UNDP country-level solutions, leveraging cutting-edge digital technology through a funding and mentorship programme.[14] Organizations can replicate and adapt these existing global initiatives, strategies and frameworks to strengthen their capacities. By creating frameworks that give practical methods to put established principles for ethical use of data into action, confidence can be created to undertake pilot projects that are monitored tohelp the organization learn and adapt to become more proficient in using data effectively.

Organizational strategy check-list

- **The creation of guidance** through written products, training programmes and/or regular webinars and workshops on such topics as transparent and consultative methodology development, ethical AI practices, tool and stakeholder frameworks for online monitoring, the nature of VE use of the internet in a given context, methodologies available for online monitoring, civil society organization/ non-governmental organization (NGO) partners available and national regulatory landscapes;
- **Collected data storage, tools and usage** to encourage in-house capacity;
- **Collective agreements on data sharing/tool usage** with technology companies following due diligence and risk assessment to reduce the burden on staff;
- **Flexible funding**, which encourages responsible piloting with the relevant learning collection processes;
- **Partner guidance and due diligence**, which include the need to understand the methodology of organizations partnered with, cost-benefit analysis and assessment of potential reputational risk.

14 UNDP (n.d.). We help digital ideas leap across borders. Available here.

CONSIDERATION 3:

Carrying out a risk assessment

How can practitioners build on risk assessment and mitigation tools to ensure that projects do no harm?

It is difficult to form a 'one-size-fits-all' static solution due to varying contexts and rapidly evolving technological and policy landscapes. Nevertheless, much has been learned about best practices for risk-aware approaches and risk assessment. UNDP has invested in creating sound practice in this area by developing data principles that are aimed to structure responsible organization-wide policy and practice when using data for development. The Principles for Digital Development can be similarly used as a foundation to form risk-informed approaches.[15] Furthermore, to safeguard a conflict-sensitive and human rights-based approach to PVE programming, UNDP requires programmes to not only integrate human rights principles and standards, but also to be informed by conflict and human rights analysis and risk assessments. UNDP has therefore developed guidance to manage risks,[16] guidance on conflict sensitivity, and a toolkit for design, monitoring and evaluation of PVE interventions.[17]

Practitioners can use their own understanding of risks ascertained from pilot projects, together with current guidelines and codes of practice for undertaking ethically informed PVE projects generally, and internet research related to VE use of the internet to create tailored risk assessments. Useful examples of resources related to specifically internet research-based risks include the ethics guidelines of the Association of Internet Researchers, the British Psychological Association's *Ethics Guidelines for Internet-mediated Research, the Norwegian National Research Ethics Committee's A Guide to Internet Research and Data & Society's Best Practises for Conducting Risky Research and Protecting Yourself from Online*

15 Lim, C. (2021). Home – Principles for Digital Development, Available here.
16 UNDP and Oslo Governance Centre (2019). Risk management for preventing violent extremism (PVE) programme – Guidance note for practitioners. Available here.
17 UNDP (2021). M&E for Preventing Violent Extremism (PVE) – UNDP Resources. Available here.
18 United Nations (2019). UN Strategy and Plan of Action on Hate Speech. Available here.

The building blocks of UNDP's risk assessment in relation to hate speech and online data

Before any policy or programming-based efforts are made to prevent and count violent extremism (VE) narratives and address hate speech, the United Nations and Member States must follow the overarching principles of international human rights law as enshrined in international instruments. These instruments include the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights, as well as the United Nations Guiding Principles on Business and Human Rights and the Rabat Plan of Action. These principles should be considered by all stakeholders and translated by organizations into actionable frameworks for using data in a human rights-compliant manner. Examples of operationalizing broader human rights law are key frameworks developed the United Nations to help international policymakers and practitioners navigate the application of human rights in this area. In addition, the United Nations Strategy and Plan of Action on Hate Speech[19] sets out the objectives of enhancing United Nations efforts to address root causes and drivers of hate speech and enabling effective responses to its impact on societies. This Plan created a clear strategy and platform to meet these objectives, which helped create clarity on the most relevant stakeholders to convene in order to reach them. The Secretary- General's United Nations Data Strategy* similarly reacts to the need for a robust, integrated approach to collecting, sorting, and using data with a plan to build technical capacity and coordination, and forming a platform to collaboratively design and implement data policies that advance the responsible human-rights-based use of data.

Note:
* United Nations Strategy-General. Data Strategy of the Secretary-General for Action by Everyone, Everywhere with Insight, Impact and Integrity 2020–22. Available here.

Harassment.[19] These guidelines can help organizations design core ethical approaches in online VE research-based work. This leaves room for adaptation in a field where the fast changing nature of the online ecosystem can quickly give rise to the need to monitor new types of online spaces, data collection and other tools and methods, research topics, etc. as determined through an approach and mind-set of M&E of projects in order to quickly adapt and react to new types of challenges and risks.[20]

Clearly defined methodologies (i.e. how data is being collected and organized, on what platforms and for what purpose) determined through an inclusive consultative process (including partners), including project risks, mitigation strategies and limitations, can go far in forming creative inclusive projects that meet their objectives.

An introduction to potential risks to consider for online data collection, analysis and application

The mind-set of risk management

"Rather than a one-off exercise, risk management is encouraged as an integral part of programming, to be carried out prior to programme design and during programme implementation." As underscored by the 2019 UNDP Risk Management Policy, at the heart of the risk management is a shift from risk aversion to responsible risk-taking. What is required for effective risk management is three-fold:

- "First, a mind-set shift: from being 'risk averse' to being risk-aware and risk-ready;
- Second, a behavioural shift: from treating risk management as separate from programming, to integrating it into the day-to-day essence of our work; and
- Third, an awareness that risk management involves identifying deviations from an expected outcome of a both negative and positive nature. Consequently, the process involves identifying both risks and opportunities associated with programming."

Note:

* United Nations Development Programme. Risk Management for Preventing Violent Extremism (PVE) Programmes. Guidance Note for Practitioners. Available here, p.12; Managing Risks Across UNDP Programming and Operations. Available here.

Key measures to mitigate potential risks:

Protect PVE practitioners and partners collecting and sharing the data, including ensuring their physical and mental wellbeing.

Protect 'at risk' individuals and communities from excessive and unwarranted surveillance by minimising the collection, storage, or dissemination of personally identifiable information (PII).

Mitigate contextual, programmatic and institutional risks through a context-specific, conflict-sensitive 'do no harm' and human rights-based approacha.

Work with partners and technical experts to build local capacity to assess risks, including potential bias of data collected and analysed.

Analyse relevant national laws to predetermine any risks regarding violations of the right to effective remedy due to data preservation policies.

Assess the level of detail of the findings by government or VE actors to be shared and their potential misuse through regular review.

19 Franzke, and others (2021). *Internet Research: Ethical Guidelines 3.0; Markham, Annette, and Elizabeth Buchanan (2012). Ethical Decision-Making and Internet Research (Version 2.0) AoIR; Ess, C. (2002). Ethical Decision-making and Internet Research*. AoIR. Links to all versions of the guidelines are available on the AoIR website at https://aoir.org/ethics; British Psychological Society (2017). *Ethics Guidelines for Internet-mediated Research*. Leicester; National Committee for Research Ethics in the Social Sciences and the Humanities (NESH) (2019). *A Guide to Internet Research Ethics*. Oslo. For listings of additional relevant guidelines, see Franzke, A.S. and others (2019) . *Internet Research: Ethical Guidelines* 3.0, 12–14; Massanari, A.L. (2018). Rethinking Research Ethics, 7; Samuel G., Gemma E. Derrick, and Thed van Leeuwen (2019). The Ethics Ecosystem: Personal Ethics, Network Governance and Regulating Actors Governing the Use of Social Media Research Data. Minerva 57, no. 3,324; Marwick, A., Lindsay Blackwell, and Katherine Lo (2016). *Best Practices for Conducting Risky Research and Protecting Yourself from Online Harassment*.New York: Data & Society Research Institute.

20 Conway, M. (2021). Online Extremism and Terrorism Research Ethics: Researcher Safety, Informed Consent, and the Need for Tailored Guidelines, Terrorism and Political Violence, 33:2, 367–380.

a) Ensuring safety measures for practitioners and targeted audiences (do no harm)

Any project that uses online data must assess any potential harm to actors involved in the data cycle, including practitioners, audiences monitored and end-beneficiary' audiences of projects. It is important to recall that online data collection and application is only one possible approach to better understand and ultimately prevent the drivers of VE.

Practitioners

The 2020 iteration of the Association of Internet Researchers' ethics guidelines raises "the growing need for protecting the researchers, as well as our subjects and informants". Welfare encompasses both physical and mental wellbeing. Being conscious of the nature of the need for resources to support the wellbeing of practitioners and access to spaces where issues can be discussed and received openly is crucial to any project that monitors potentially disturbing material. When sharing research also, potential ramifications on the research source if not anonymized, should be included in risk management processes. "...Additional knowledge generation around the issues, dedicated resources to mitigate some of the potential risks and harms, and increased training for all researchers..." who are active in the sub-field of VE-related research online is needed to understand and mitigate some of the risks to wellbeing posed by these types of projects.[21]

Audiences targeted online by monitoring

The United Nations General Assembly Resolution 68/167 on the right to privacy in the digital age describes the "unlawful or arbitrary collection of personal data" as a highly intrusive act that could violate "the rights to privacy and to freedom of expression and may contradict the tenets of a democratic society".[22] The right to privacy protects each individual's "private sphere", an "area of autonomous development, interaction and liberty" where they are safeguarded "from state intervention and from excessive unsolicited intervention by other uninvited individuals".[23] While national privacy laws and frameworks vary in content, most follow a set of common principles, including that personal data processing should be "fair, lawful and transparent", as well as limited to what is "necessary and proportionate to a legitimate aim".[24]

One UNDP project applies a public health-based surveillance model for the collection, analysis and presentation of social media analytics that focuses on risk factors rather than profiling specific communities and individuals. The right to non-discrimination guarantees that no person shall be treated less favourably because they hold certain protected characteristics such as race, gender, ethnic origin, or religion.[25] Through this approach, UNDP aims to promote the protection of at-risk individuals and communities from excessive and unwarranted surveillance. Therefore, efforts should be made to minimize the collection, storage, and dissemination of personally identifiable information (PII), except where it is warranted, and/or material to the findings of the study. Consequently, UNDP and its project partner only uses publicly accessible social media data procured and/or accessed within the relevant privacy and terms of service requirements. This use of publicly made data is supported, for example, by Article 9.2(e) of the European Union's General Data Protection Regulation (GDPR), allowing for the processing of what it terms "special categories of personal data" if it has been "manifestly made public by the data subject". Research is only engaged in compliance with the national law of the partner's base country and United Nations principles, and in particular the Universal Declaration of Human Rights, and ethical codes of conduct. Effort above and beyond reliance on human rights frameworks is needed in risk assessment, particularly since a complex picture of what freedom of speech represents in the online sphere today could lead to inadequate protection of vulnerable persons monitored online in various jurisdictions. Tailored risk assessment measures, particularly when training AI models include personal data encoding in manual training of the AI technology and mitigation of personal political bias in treating the information

Audience(s) targeted by data applied through programmes and projects

The projects and programmes that are informed by applying the online data collected and analysed by PVE practitioners must ultimately undergo a risk assessment to ensure that those impacted by such activities are protected by systems in place. For example,

21 Conway, M. (2021). Online Extremism and Terrorism Research Ethics: Researcher Safety, Informed Consent, and the Need for Tailored Guidelines, Terrorism and Political Violence, 33:2, 367–380; Allam, H. (2019). It Gets to You. Extremism Researchers Confront the Unseen Toll of Their Work. NPR, 20 September 2019; Berger, J. M. (2019). Researching Violent Extremism: The State of Play. Washington DC: RESOLVE Network; Krona,K. (2020). Vicarious Trauma from Online Extremism Research: A Call to Action. GNET Insights, 27 March 2020.

22 United Nations General Assembly (21 January 2014). Resolution adopted by the General Assembly on 18 December 2013. 68/167.

23 United Nations General Assembly (17 April 2013). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue. A/HRC/23/40. Available here.

24 Human Rights Council (3 August 2018). The right to privacy in the digital age: Report of the United Nations High Commissioner for Human Rights. A/HRC/39/29. Available here.

25 United Nations and the Rule of Law, Equality and Non-discrimination. Available here.

the Guidance Note on Risk Management for PVE programmes[26] and the supplementary note to support revision of risk management strategies in the context of COVID-19[27] are designed – in simple, non-technical terms – to enhance practitioners' understanding of risk management as an approach when working in complex settings. Simultaneously, it brings attention to the contextual, programmatic and institutional risks associated with working on PVE programmes specifically, drawing attention to how a context-specific, conflict-sensitive 'do no harm' and human rights-based approach can help to mitigate many of these risks and improve the effectiveness and efficiency of PVE programmes.

b) Building capacity of partners to assess risks

Central to risk mitigation relating to online-based projects are issues of privacy, personal data protection and freedom of expression. However, particularly in post-authoritarian contexts, since there is often a lack of specialized knowledge and understanding of these aspects of working with data, building capacity of local partners and even national staff members is a challenge.

UNDP has built national capacity by issuing calls for contracts for international firms that specialize in monitoring to work with NGOs/CSOs, for the mutual benefit of all partners to enhance both technical and contextual expertise. Utilizing and building capacity of actors with greater contextual understanding can help limit the risks associated with bias and lack of understanding of nuance and language in data collected (*see Considerations 5* and *7* for a summary of the types of risks to be taken into consideration related to the accuracy of data collected in the data cycle). However, undertaking this can often also pose significant security risks, especially in situations of political unrest where potential risks to partners or groups monitored may make capacity building unsafe. An approach of do no harm must always lead projects.

c) Data preservation

Incorporated into these risk assessments should also be the storage/preservation of data. National laws may place timed limitations on data storage. However, Article 8 of the Universal Declaration of Human Rights (UDHR) states that "everyone has the right to an effective remedy by the competent national tribunals for acts violating the fundamental rights granted him by the constitution or by law," with the United Nations Guiding Principles on Business and Human Rights stating that companies should "provide for or cooperate in their remediation through legitimate processes" when associated with adverse human rights impacts. Therefore, national and international law must be analysed through risk assessment to predetermine whether or not the ability of affected rights holders to realize their right to effective remedy may be compromised if content collected through these projects is deleted and no longer available for use in various scenarios, such as investigating serious crimes under international humanitarian law.

d) Sharing research

The securitization of preventing and countering violent extremism (P/CVE) in many contexts poses a challenge for development actors to publish research on issues pertaining to PVE. Insights by international organizations must be provided with high sensitivity since international reputation is at risk. The stance of and commentary by international actors on VE can be seen as giving political leverage to VE actors who attempt to discredit the government's prevention and containment efforts by linking them with international aid and foreign pressure. While public launch and dissemination of findings can be inadvisable due to risks such as these, insights from analyses can still be used to inform internal as well as partners' peacebuilding policy and programming such as those focused on building resilience to VE messaging through education and messaging.

However, sharing findings with government actors, acquired from online monitoring, can encourage further securitization by using them to legitimize the reduction of freedom of expression, the illegal use/breaching of privacy of personal data and/or worsening human rights either online or offline. Many governments and communities are sensitive to the idea that violent extremists are active in their country or in their cyberspace. If not managed in an effective and coordinated way, such sensitive information could adversely affect wider PVE programmes of an organization and the wider community that an organization is interested in protecting. Additionally, the support and reputation of bodies undertaking monitoring should not be instrumentalized by law enforcement and security agencies to further securitize the national P/CVE space. Therefore, a risk mitigation strategy for the communication and dissemination of findings must be put in place.

26 UNDP Oslo Governance Centre (2019).Risk Management for Preventing Violent Extremism (PVE) Programmes Guidance Note for Practitioners. Available here.
27 Risk Management for Preventing Violent Extremism (PVE) Programmes – Guidance Note for Practitioners: Supplementary note to support revision of risk management strategies in context of COVID-19, Available here.

Another consideration when sharing results of data are that agency and donor logos on information and communication material used in PVE projects should at times be avoided if they are perceived to do harm through risk assessment. Rather, it could be more effective to align findings to governments' ongoing work and integrate them into reports and analysis in this regard; outputs can also be communicated independently from their source of origin, i.e. the United Nations entity, the organization and/or donors.

The level of risk regarding the misuse of findings by government or VE actors can be assessed against regular review of news and secondary information to keep abreast of policies that may heighten the relevance of research findings. This can be complemented, if possible, with regular contact with government counterparts to understand their perspectives on peacebuilding and PVE interventions, and sensitivities surrounding certain related issues. Employing staff with national political understanding and/or experience managing sensitive political files can also help inform this risk management process.

The detail of research shared, including concerning the actors identified as extremists or as influential among them, should also be assessed in addition to the risk that national media coverage may sensationalize, politicize and omit nuance or give a platform to misrepresenting facts.[28] Steps should be taken to ensure that at-risk individuals and communities are protected from excessive and unwarranted surveillance, and that no PII that can be misused to target vulnerable groups is shared.

The importance of language when sharing research

Preventing violent extremism (PVE) practitioners can utilize online data to understand trends and address factors contributing to radicalization that leads to violent extremism (VE), rather than radicalization more broadly, since radicalization can – and is most likely to – occur without violence. PVE practitioners' concern is the prevention of VE and hate speech leading to violence. Language is important, particularly when sharing findings of online-based research as misuse of language can create opportunities for undue persecution of actors (or those associated with actors) highlighted.

> "There are also risks of human rights violations when the terms 'extremism' or 'radicalization' are used to cover non-violent activity. States should ensure that the focus of their measures is on actual conduct, rather than mere opinions or beliefs. International human rights law provides a clear framework for the promotion and protection of human rights. In particular, the right to hold an opinion and the freedom to have or adopt a religion or belief of one's choice cannot be subject to any restrictions."*

Source: United Nations Office of Counter-Terrorism (2019). *Developing National and Regional Action Plans to Prevent Violent Extremism*, p. 23.

28 Berger, J.M. (2019). Researching Violent Extremism: The State of Play. Washington, D.C.: RESOLVE Network.

CONSIDERATION 4:

Encouraging multi-stakeholder approaches that do no harm (due diligence of partnerships)

How can practitioners use the opportunities provided by partners' expertise while ensuring that processes do no harm?

By engaging partners including civil society and tech platforms, new expertise, technology and capacity can be leveraged to enhance quality, efficiency, legitimacy and relevance of interventions. Partners can be selected to help overcome challenges identified at the design stage. However, obtaining partnerships should never be seen as a shifting of responsibility. Rigorous verification of stakeholder ethical and human-rights standards must come first and foremost in partner consideration. Part of this entails ensuring that stated methodologies are transparent, as are employment standards of those working for the organization in question, and data collection and storage practices are systematically risk- assessed and human rights-compliant. Funding towards partnerships must be justified against the project objectives, ensuring both quality and value assurance. The nature of associated opportunities, risks and responsibilities can differ depending on the scale of the partner utilized.

The United Nations Strategy and Plan of Action on Hate Speech was developed on the basis of a joint effort by 14 United Nations entities, and tasks the United Nations with addressing "the root causes and drivers of hate speech", on the one hand, and enabling effective responses to its impact upon societies, on the other. [29] A subsequent Guidance was developed by the United Nations Office on Genocide Prevention and the Responsibility to Protect, the designated United Nations focal point on the Strategy, to provide more detailed advice and direction on how the Strategy should be effectively implemented by United Nations field presences.[30]

29 United Nations (2019). United Nations Strategy and Plan of Action on Hate Speech. *United Nations Report, May*. Available here.

30 United Nations. (2020). *United Nations Strategy and Plan of Action on Hate Speech Detailed Guidance on Implementation for United Nations Field Presences.*

Action 17: Engaging private actors

Commitment 6 of the Guidance is using technology to combat hate speech:

United Nations entities should keep up with technological innovation and encourage more research on the relationship between the misuse of the Internet and social media for spreading hate speech and the factors that drive individuals towards violence. United Nations entities should also engage private actors, including social media companies, on steps that they can take to support United Nations principles and action to address and counter hate speech, encouraging partnership between government, industry and civil society.

Action 17 under this commitment provides guidance on engaging private actors when using technology to combat hate speech as seen in the extract from the Guidance below.

67) *Identify which tech and social media companies are most relevant in the particular country context and prioritize building partnerships on the implementation of the Strategy accordingly, especially in relation to the monitoring and collection of data on hate speech.*

68) *Through their interactions with representatives of social media platforms, and bearing in mind the impartiality of the United Nations, encourage these companies to:*

68.1) *Respect human rights as required under the Guiding Principles on Business and Human Rights and, in doing so, evaluate how their products and services affect the human rights of their users and public, through periodic and publicly available human rights impact assessments;*

68.2) *Align their content policies on hate speech with international human rights norms and standards, including the Rabat Plan of Action.*

Stakeholder snapshot – Utilizing specific expertise through partnerships

Partners can be selected to specifically overcome challenges such as:

- ***Capitalizing on expertise to overcome algorithmic biases***
  For example, the **Citibeats**[i] platform follows an ethics-first methodology for AI data collection with the aim to transform people's opinion into actionable data. Algorithms utilize machine learning technology such as natural language processing, and analysis is driven by respect of privacy, bias removal (referred to the data, the training process of the AI, the extraction of insights, and the interpretation of those insights).

- ***Capitalizing on expertise to overcome data privacy concerns and reaching 'hard to reach' regions for sensitive issues***
  For example, **RIWI**[ii] is a global survey technology and sentiment analysis firm that gathers citizen opinion data by inviting randomized Web users in every geo-targeted area in the world to anonymously participate in a language-appropriate survey or specific citizen engagement initiative.

- ***Capitalizing on expert to directly link and monitor online interventions with offline impact***
  For example, **Moonshot**[iii] targets online users who may be more vulnerable to violent extremism (VE) content with advertising encouraging mindfulness and links to local service providers. They train mental health providers with VE experts in order to understand the most effective methods of helping build resilience.

- ***Capitalizing on local expertise to more accurately collect data in a human rights-compliant manner***
  For example, **Koe Koe Tech**[iv] has developed a platform for CSOs to coordinate flagged harmful content, guided by the principle of an international human rights law, particularly ICCPR, to contextualise and hold big tech companies to account in their content moderation processes.

- ***Capitalizing on lexicons to tailor to particular contexts in order to reduce burdens of developing methodologies from scratch***
  For example, **PeaceTech Lab**'s[v] series of hate speech lexicons identify and explain inflammatory language on social media while offering alternative words and phrases that can be used to combat the spread of hate speech in conflict-affected countries.

Notes:

i Citibeats (2021) Citibeats: Understanding What Matters to People, at Scale in Real-Time. Available here.
ii 'RIWI | Global Trend-Tracking and Prediction Technology' (*RIWI*, 2021). Available here.
iii Moonshot (2021). Available here.
iv 'Koe Koe Tech' (2021). Available here.
v Peacetech Lab (2021) Our Hate Speech Lexicons | Peacetech Lab' *Putting the Right Tools in the Right Hands to Build Peace*, 2021). Available here.

Civil society organizations should be prioritized as partners due to the contextual expertise offered. Much discussion online is highly context-specific and therefore needs local knowledge and, fluency in local languages and dialects. Granular local knowledge can help mitigate some of the biases amplified by machine learning algorithms.

UNDP Sri Lanka working closely with the Resident Coordinator's Office has utilized the partner Hashtag Generation for their work in monitoring dangerous speech online in the country. Hashtag Generation is a movement led and run by a group of young Sri Lankans advocating for the meaningful civic and political participation of youth, especially young women and young people from minority groups.* The use of a local partner able to mobilize dedicated resources and sufficient vernacular language for data cleaning and generating narrative reports has been both useful and cost-effective for the United Nations in Sri Lanka.** The United Nations has also been able to provide initial technical support to Hashtag Generation in setting up monitoring approaches through the Peace and Development team.

- Hashtag Generation is a movement led and run by a group of young tech-savvy, socially conscious Sri Lankans advocating for the meaningful civic and political participation of youth, especially young women and young people from minority groups. The group adopts a non-partisan approach and works with the strong conviction that decision-making at all levels should remain transparent and inclusive and in order to remain sustainable and build lasting peace in Sri Lanka.

The UNDP Office in Sri Lanka commented that the sorting, data cleaning and generating user-friendly narrative reports is very time-consuming and the use of a local partner able to mobilize dedicated resources and sufficient vernacular language staff has been both useful and cost effective for UNDP Sri Lanka.

Notes:

*Hashtag Generation (2021).'About - Hashtag Generation. Available here.

** ibid.

Through this project, keyword searches on the platforms used aim to understand the levels of reach/engagement that harmful narratives receive. Monitoring via Crowdtangle is supplemented with searches based on keywords from a comprehensive lexicon (e.g. of slurs) as well as proactively monitoring actors of concern. The wide categories of pages and public groups monitored includes less conventional but highly popular sources of social media based information such as gossip news pages and meme pages. These pages were selected based on two criterions: pages that have high levels of reach among Sri Lankan internet users and pages that have a history of circulating and disseminating dangerous speech narratives. Currently over 1,100 pages and 100 groups are monitored. The data captured are disaggregated where possible, based on the demographic data of the type of actor engaging in this type of speech (where this is publicly available). Where there are clear attacks aimed at a specific individual or a community, this data is also recorded.

The combination of tooling and local expertise has allowed the project to utilize trilingual (Sinhala, Tamil and English) capability to identify content that amounts to dangerous speech including hate speech and disinformation. A central finding of this project has been the large impact of current affairs on dangerous speech and the way VE groups aim to utilize these day-to-day narratives to enhance their more long-term narratives that ultimately aim to contribute to the 'us vs. them' mentality.

Private companies specializing in monitoring

BENEFIT:

By engaging partners with monitoring expertise, technology and capacity can be enhanced to increase the quality, efficiency, legitimacy and relevance of interventions. Professional organizations will also have experience in the risks and mitigation strategies of projects that access and utilize online data, as well as in communicating findings in an impactful manner. These organizations will likely also have an understanding of the best ways to access data from the big platforms through experience and established relationships.

CONSIDERATION:

Ensuring that stated methodologies and associated costs are transparent (i.e. exactly how objectives will be reached and communicated), as are employment standards of those working for the organization in question, and data collection and storage practices are systematically risk assessed and human rights compliant. Funding to establish and maintain partnerships should be justified against the project objectives, ensuring both quality and value assurance.

Civil society organization/non-governmental organization partners

BENEFIT:

Much discussion online is highly context-specific and therefore needs local knowledge, and fluency in local languages and dialects. This is particularly beneficial considering the locally relevant, dynamic, and constantly evolving nature of language and dialects. Granular local knowledge can help mitigate some of the biases amplified by machine-learning algorithms. CSOs can also go far in validating data gathered online, which is generally more difficult to achieve than data gathered through traditional collection methods such as interviews.

CONSIDERATION:

CSO's rights and wellbeing must be preserved through any PVE project that utilizes online data and AI. Those who work in the highly sensitive area of PVE can be targeted both online and offline, and risk exposure to upsetting material.

Major tech platforms

BENEFIT:

Access to data is a key component in measuring and tracking violent extremist content and sentiment that is propagated on and through the internet. Large tech companies often serve as gatekeepers of data needed to train and develop AI algorithms if tools are being developed in-house, and usually have a specialized skills and higher capacity to design, develop and maintain innovative technological tools to monitor VE trends.

CONSIDERATION:

Transparency on human rights compliance of these platforms' terms of service, including user consent for data usage, and redress mechanisms, is needed when engaging in PVE. Additionally, meaningful external, including democratic oversight, is a crucial consideration when using the data of such platforms.

The bargaining power of practitioners vis-à-vis big tech companies can be limited.

Big tech business models

- Patterns of hate speech, dehumanization and identity-based narratives have been demonstrated as contributing to conditions where violent extremism (VE) becomes more likely. Society can be polarized through algorithms that encourage users to view agreeable information encouraging individuals to feed off different facts from each other.[i]
- Research suggests that more extreme content closer to platform terms of service (i.e. the terms and conditions of platform use), will get more views.[ii] This is compounded by the fact that, due to this rise, recommender systems may promote extreme content.[iii]
- The argument that recommendation algorithms help create VE online is not clear due to questions on the extent and true impact of these systems.[iv] However, on the most harmful end of the scale is the fact that algorithms have been shown to have the potential to increasingly feed – and indeed have fed – an initial interest in extreme material,[v] aiding the creation of alternative news networks.[vi] Algorithms have become more apt at finding 'rabbit holes'[vii] or 'filter bubbles' for individuals to get lost in online and bypass thoughtful consideration by dramatically amplifying confirmation bias.[viii]
- This occurs by the use of positive intermittent reinforcement techniques to manipulate dopamine release, in order to keep users engaged online and therefore more likely to come into contact with advertisements. This constant stream of information encourages 'system 1 thinking', conducive to violent extremist groups, operating 'automatically and quickly, with little or no effort and no sense of voluntary control'.[ix]
- Since AI-enabled technologies have the potential to influence individuals' thoughts, there is a clear relevance to the right to freedom of thought in its internal dimension.[x]

Notes:

[i] HOPE not hate, 'State of Hate 2020: Far Right Terror Goes Global' (2020). Available here.

[ii] Zuckerberg, M. (2018). *A Blueprint for Content Governance and Enforcement.*

[iii] Whittaker, J., S. Looney, A. Reed, and F. Votta (2021). Recommender systems and the amplification of extremist content. *Internet Policy Review, 10*(2). Available here;
O'Callaghan, D., D. Greene, M. Conway, J. Carthy, and P. Cunningham (2015). Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems. *Social Science Computer Review, 33(4), 459–478; Council of the European Union (2020). The role of algorithmic amplification in promoting violent and extremist content and its dissemination on platforms and social media; Tech Against Terrorism (2021). Position paper Content personalisation and the online dissemination of terrorist and violent extremist content.*

[iv] Ledwich, M., and A. Zaitsev (2020). Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *First Monday* (suggests that effects may be limited); Chen, A. Y., B. Nyhan, R.R. Robertson, J. Reifler, and C. Wilson (2020). *Exposure to Alternative & Extremist Content on YouTube* (suggests individuals that are recommended extreme content already have extreme views).

[v] Reed, A. and others (2019). Radical Filter Bubbles: Social Media Personalisation Algorithms and Extremist Content. *Global Research Network on Terrorism and Technology*: Paper No. 8.

[vi] Lewis,R. (2018). Alternative Influence: Broadcasting the Reactionary Right On Youtube. *Data & Society.*

[vii] O'Callaghan, D., and others (2014). 'Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems. Social Science Computer Review, 33; Cain, C. (2019). Caleb Cain Was a College Dropout Looking for Direction. He Turned to Youtube. *The New York Times.*

[viii] Pariser, E. (2013). *The Filter Bubble: What the Internet is Hiding from You.* Penguin.

[ix] Ingram, H. (2016). Deciphering the Siren Call of Militant Islamist Propaganda: Meaning, Credibility & Behavioural Change. *Terrorism and Counter-Terrorism Studies.*

[x] Human Rights Committee. (27 September 1993). General comment No. 22 (48) (Art. 18). CCPR/C/21/Rev.1/Add.4. Available here; Allegre, S. (2017). Opinion: Rethinking Freedom of Thought for the 21st Century. *European Human Rights Law Review*, 3, 224.

Consideration of whether the business model of partners is in fact at odds with efforts to prevent divisive, harmful and violent content should be incorporated into risk assessment.

Partnership with development actors can act as a positive guise of human rights compliance for tech companies; hence, any partnership or agreement should take this into consideration and follow due diligence, which can encourage positive, mutual working relationships.

The "Human Rights Due Diligence Training Facilitation Guide" provides flexible training modules that clarify what is required for companies to conduct human rights due diligence.[31] The Guide is complemented by a Human Rights Self-Assessment Training Tool featuring 99 potential business-related human rights risks

31 UNDP (2021). Human Rights Due Diligence Training Facilitation Guide. Available here.

with references to international human rights instruments and relevant SDGs.[32] In addition, in order to help guide United Nations staff in undertaking due diligence within partner selection, the United Nations advocates for key messaging when engaging with potential partners in order to ensure that an awareness of the human rights implications of stakeholder practice and policy frames all potential collaboration.

Tailored due diligence:

UNDP was interested in partnering with a major tech/social media company as part of a programme to build the skills of young influencers to disseminate positive messages to tackle hate speech and promote tolerance for diversity online. Given the controversies surrounding the company, particularly around privacy protection of users and alleged lack of censorship of pornographic content, the implementing project was advised by the partnership unit to conduct a robust due diligence. This process, including the development of a Risk Assessment Tool, Risk Log, Risk Monitoring Plan and Strategic and Crisis Communications Plan, aims to prepare the project and UNDP as a whole to effectively mitigate the reputational and policy risks that might be associated with such partnerships. While controversial cases of human rights and environmental abuse are not new to any tech giants, implementing effective risk mitigation strategies would allow for UNDP to form new partnerships and still protect itself from allegations of harm associated with the partnering company. Due diligence is critical for financial and non-financial partnership with private sector companies in order for the project to be well prepared for any risks associated with the engagement, and it should be considered a prerequisite for the signing of any agreements (memoranda of understanding, letters of understanding, etc.) with these firms.

32 UNDP (2021). Human Rights Due Diligence Training Facilitation Guide. Available here.

@UNDP Kyrgyzstan

II

COLLECTION, ANALYSIS AND APPLICATION OF ONLINE DATA

CONSIDERATION 5:
Accessing relevant data

What data is available to practitioners to monitor for PVE and how has policy changed the nature of data available?

The data made available to practitioners is impacted by the policy and practice of regional, national and private actors, which can be made in response to public pressure pertaining to security, privacy and/or freedom of expression, factors that are often competing. The policy landscape in which violent extremist use of the internet is regulated and data monitored by external actors is therefore complex and compounded by the global reach of private companies. Violent extremist actors often respond to and manoeuvre policy, which then further changes the nature of the data related to VE online, including the most prominent data sources (online platforms). In order to create data-driven projects that set realistic and achievable aims, the data available to practitioners must be understood in a given national context.[33]

Digital research can provide insight on the gender, demographic profile, location and other measurable characteristics of followers when they consent to sharing this data on public platforms and using this data is authorized by the Terms of Service of the platform(s) in question, unless other data sharing agreements have been reached. If this information is not shared by users or made accessible by platforms, it is possible to glean insights from information available, for example, the location of social media commentary could be generally inferred from language or dialect used, subject matter, timing of posts, etc. Human creativity and flexibility in data collection is therefore necessary in order to reach project objectives meaningfully despite limitations that will inevitably be faced regarding content available online. The nature of online data made available from social media platforms is the information that individuals choose to share. Material intentionally shared for public display is a purposefully constructed, digitally mediated identity, or digital avatar.

Material that incites violence is now often removed automatically by platform content removal algorithms, and platforms have also increased privacy considerations when sharing data for research or commercial considerations due to increased political and public pressure. A reduction in available data has been accelerated particularly by two events – the live streaming and subsequent posting of the Christchurch massacre of 2019, leading to greater regulatory policy and practice, and the Facebook Cambridge Analytica scandal, leading to heightened user privacy considerations. The latter, reported first in 2015, triggered a more 'privacy-focused' future for Facebook, with Mark Zuckerberg claiming that the incident had 'fundamentally altered [its] DNA'. Clearly, heightened user privacy is ultimately a positive development from a risk assessment perspective, but PVE practitioners should be aware of the limits of this data for PVE, as well as consider the other risks that are still active when using the data and resources of large tech platforms.

To overcome some of these issues, Tech Against Terrorism, a public-private partnership supported by the United Nations Counter-Terrorism Committee Executive Directorate, built the Terrorist Content Analytics Platform (TCAP), which alerts tech companies with terrorist content when found on their platforms. With renewed funding, TCAP will expand its archiving function, which will keep record of terrorist content prior to tech companies removing it from their platforms. This content will be made available to academics and civil society practitioners to support evidence-based research that guides P/CVE policies.

Regulatory efforts to combat VE utilization of the internet

The multi-layered policy landscape that has evolved following the threat of VE utilization of the internet is a space where responsibility is often passed between private and governmental

33 For a breakdown of data available by platform, see Build Up's Social Media Analysis Toolkit, 'Understanding what data is available'. Available here.

actors, creating confusion, and from this confusion, a lack of clear lines of responsibility and transparency of guiding principles, rules and regulations. This is a danger where the implications of policy made by these actors impacts the whole of society. Clarity, understanding and transparency, then, are key to manoeuvring this space, with an acknowledgement that each actor, policy and practice is part of the solution and therefore must be understood in complementarity.

CVE practices of content removal necessarily impact the data available to PVE practitioners on VE actors, audiences and content online, and are therefore important to include in any discussion of policy frameworks in this area. Different actors have attempted to reduce the strategic worth of the internet for VE groups.

The United Nations Plan of Action to Prevent Violent Extremism (2015) states:

Violent extremism is a diverse phenomenon, without clear definition. It is neither new nor exclusive to any region, nationality or system of belief. Definitions of "terrorism" and "violent extremism" are the prerogative of Member States and must be consistent with their obligations under international law, in particular international human rights law.

Source: United Nations (2015). Plan of Action to Prevent Violent Extremism. New York.

The large platforms

Prior to the development of national and regional legislation on regulation, social media companies effectively had to self-regulate. Efforts were made to introduce and enforce regulatory standards; however, the need emerged to define the difference between illegal and harmful, yet legal content. This is particularly pertinent in the ever-complex landscape of VE rhetoric online, compounded by competing issues of contextual nuance and algorithmic processes scaling up removal efforts. Larger companies that can afford safety and security teams now make extensive use of AI in their efforts to remove and block terrorist content from their platforms.[34] Since these efforts aim to channel human conduct, transparent content moderation, then, is hugely important, with accessible and comprehensive appeals processes necessary in order to understand their legal basis. Responses to VE use of platforms – including 'deplatforming', i.e. preventing the use of a platform; and 'shadow-banning', i.e. deletion or limiting visibility of content, without the user being aware – can potentially lead to subjective, disparate and potentially biased enforcement, which results partly from the lack of internationally agreed on definitions of what constitutes terrorism and VE.[35] This can lead to inaccuracy and bias in results of monitoring VE-related activity online.

Large technology companies utilize classifiers to remove dangerous content, which greatly reduces the posting and reposting of illegal content. However, VE groups continuously attempt to outmanoeuvre classifiers with subtle changes, and new events create large flows of content that are difficult to detect in a timely manner. In addition, emerging platforms that gain popularity quickly often fail to keep up with threats posed by those aiming to utilize such reach.[36] Despite technology companies having enormous power of regulation, blaming them for failures in terms of VE utilization of their platforms is not constructive if not paired with solutions, and often fails to account for the fact that this responsibility is often sub-contracted to these private entities by states due to the difficulty of state regulation of such a complex phenomenon. Additionally, when errors do occur, states often do not want to or lack the technical capacity to face this problem and can struggle to legislate in a way that balances competing values of privacy, expression and security.

Governmental efforts

Resolution 75/291 calls for "due diligence to be applied by hosting service providers, in line with national legislation [...] and the Guiding Principles on Business and Human Rights, in order to address the dissemination to the public of terrorist content through their online services, including through the lawful removal of terrorist content" while recalling that "the primary responsibility to counter incitement to commit terrorist acts and to promote and protect human rights and fundamental freedoms lies with the State". Hence, regulation of digital communications must always comply with relevant requirements of international human rights law, including the need to fulfil the three-fold test of legality, necessity and proportionality whenever restricting the rights to freedom of expression or privacy. Government efforts (national

34 Macdonald, S., Sara Giro Correia, and Amy-Louise Watkin (2019). Regulating Terrorist Content on Social Media: Automation and the Rule of Law. Cambridge University Press.

35 United Nations Office of Counter-Terrorism (UNOCT) (2021). Countering Terrorism Online with Artificial Intelligence. Available here; Human Rights Watch (30 July 2020). Joint Letter to New Executive Director. Global Internet Forum to Counter Terrorism. Available here.

35 Ciarán O'Connor, C. (2021). Hatescape: An In-Depth Analysis of Extremism and Hate Speech on TikTok. ISD. Available here.

and regional) to counter VE use of the internet have rested largely on passing responsibility to social media companies to remove content deemed 'harmful' in a timely manner and heavy fines in cases of non-compliance. The major issue with tackling 'harmful' content is that it necessarily relies on a subjective interpretation of harm that is particularly difficult to implement at scale without negatively impacting on freedom of expression. Material that does not violate the law but potentially does violate platform Terms of Service leaves a grey area where discretion is generally left to the platforms themselves.[37] Tech Against Terrorism has highlighted two other main themes of governmental regulation in this area as proposals that empower removal or correction orders to platforms on the basis of mis- or dis- information, as well as laws motivated by the idea that adaptation to the digital space is necessary. This type of regulation is often seen as a package of regulation, incorporating both content moderation and user privacy considerations in response to the changes seen as a result of today's digital world.[38] Increasing use of national authorities' own removal of online VE content has also been seen by using the content flagging mechanisms in the platforms to report content as an infringement of the platforms' terms of service.[39]

Multi-sectoral cooperation

One result of the fragmented nature of this space, together with increased public pressure, has been calls for increased cooperation between these platforms and government actors, together with civil society and academia, to ensure consistent content moderation policies. The Global Internet Forum to Counter Terrorism (GIFCT), for example, has innovated in a shared hash database, allowing members to share digital fingerprints of terrorist content found in URLs and PDFs which in turn will encourage a wider breadth of detection of VE content. For URL-based hashed efforts GIFCT is working with the TCAP built by Tech Against Terrorism to begin feeding these hashed URLs into the GIFCT's hash database to support smaller tech companies with automated removals. While GIFCT and others can propose frameworks and encourage companies to develop further tooling and transparency efforts, there are still a number of companies that do not participate in these self-regulatory and multi-stakeholder discussions and settings.

Key principles for creating standards in this area that should be valued are transparency of the mechanisms and actors making standards and how accountable those standards are.[40] "When there is standard setting and regulation in an area that has such a profound effect on human life on this planet in the 21st century, we should be asking really hard questions about transparency, governance and access to those spaces."[41]

The intersection of P/CVE policies and practice (the impact of takedown methods on PVE approaches)

The United Nations Plan of Action to Prevent Violent Extremism Action Plan notes that "national plans should be developed ... to include countering and preventing violent extremism measures".[42]

The technological manoeuvring of VE groups – How data available has evolved with policy to remove terrorists from the open web

With the rapid spread of Daesh (so-called *Islamic State*) propaganda activity through videos and posts on social media and anonymous sharing platforms came a rise in terrorist activity in the Middle East and Europe. The US Government therefore established countermeasures such as takedown, account suspensions and hacker attacks against Daesh.[i] Twitter began a campaign of suspensions in response to Daesh's 'Golden Age' on the platform.[ii] To adapt to these measures, Daesh moved its propaganda network more heavily towards encrypted communication platforms and file-sharing sites.[iii] This allowed so-called *jihadi* groups community resilience in order to continue to generate content, disseminate propaganda and communicate freely within safe online havens through the use of an array of technological applications.[iv]

This move was also significant in terms of radicalization because these sites created 'black boxes' for Daesh-related propaganda where viewers could become emerged in it.[v] The term 'jihadwiki' has been used to describe the online libraries of terrorist material that was stored together, rather

37 Keller, D. (2021). If Lawmakers Don't Like Platforms' Speech Rules, Here's What They Can Do About It. Spoiler: The Options Aren't Great. Available here.
38 Tech Against Terrorism (2021). The Online Regulation Series: The Handbook. Available at shorturl.at/msBLM>; See an update of the Tech Against Terrorism handbook, including an examination of regulation in updated national settings here: Tech Against Terrorism. (2021). *THE ONLINE REGULATION SERIES 2021 | GENERAL UPDATE.*
39 For example, see Europol (2021). EU Internet Referral Unit EU IRU. Available here.
40 Douek,D. (2020). The Rise of Content Cartels. *SSRN Electronic Journal.*
41 Chatham House (n.d.) Countering Terrorist Use of the Internet. Available here.
42 United Nations (2015). *Plan of Action to Prevent Violent Extremism: Report of the Secretary-General*. 24 December. A/70/674, para. 44.

than scattered over more mainstream surface sites such as Facebook and Twitter.[vi] Much extremist material is now stored in the 'Dark Web', where content is not indexed by standard search engines and is intentionally concealed.[vii] This clearly leads to radicalized individuals being more able to evade authorities which have traditionally used the surface web to pre-empt violent acts. As a 2015 special report stated, "while the Dark Web may lack the broad appeal that is available on the Surface Web, the hidden ecosystem is conducive for propaganda, recruitment, financing and planning, which relates to our original understanding of the Dark Web as an unregulated space".[viii] However, since the Dark Web demands a relatively high barrier to entry in terms of technical ability, encrypted messaging services such as Telegram and Signal remain the more likely choice for violent extremist actors to utilize.

Telegram has become a popular platform due to its internal file-sharing capabilities. By proliferating joinlinks and observing standard cyber security measures, users can share unlimited content on a one-to-one basis or within in-groups/super groups which are either openly accessible within Telegram or by invite i.e. 'joinlink' only.[ix] 'Fanboys' have disseminated messages to a wider audience, by creating Telegram groups with multiple links and ready-made throwaway twitter accounts. A proportion of tweets have been found to contain links to news reports and coverage of Daesh material, which raises important questions about the role of traditional news media in spreading terrorist propaganda.[x] The media has been considered by violent extremist groups as an effective weapon.[xi]

Conspiracy theories and fake news have become increasingly powerful methods of extending the potential reach of VE efforts towards radicalization into the mainstream. For example, the far-right has attempted to normalize its views by attaching them to health fears.[xii] Widely ranging advocates of extremist ideologies have used COVID-19 conspiracy theories around anti-vaccination, anti-establishment, anti-minority and anti-Semitic disinformation in order to further their own ideological aims.[xiii]

As the threat landscape online adapts, content moderating AI models which are trained to filter out specific content based on certain criteria suffer from inherent limitations. For instance, a machine learning model trained to find content from one VE organization may not work for another because of language and stylistic differences in their propaganda. Complexity in language use, particularly related to humour,[xiv] and the fact that current models are often trained on major languages and are therefore less reliable for minority languages, necessitate human oversight of the review and decision-making processes despite the sheer volume of content to be monitored.[xv]

Notes:

[i] Shehabat, A., and T.E. Mitew (2018). Black-Boxing the Black Flag: Anonymous Sharing Platforms and ISIS Content Distribution Tactics. *12 Perspectives on Terrorism*, p. 81.

[ii] Aly, A. and others (2016). Introduction', *Violent extremism online: new perspectives on terrorism and the internet*. Routledge. p.3.

[iii] Conway, M. (2016). Determining the Role of the Internet in Violent Extremism and Terrorism: Six Suggestions for Progressing Research. *40 Studies in Conflict & Terrorism*. p.82.

[iv] ibid.

[v] 'Tech for Jihad: Dissecting Jihadists' Digital Toolbox' (2016). Available here.

[vi] Weimann, G. (2016). 'Terrorist Migration to the Dark Web' (2016) *10 Perspectives on Terrorism*. p. 41; SITE Intelligence Group (2014). Jihadist Suggests Creating 'Jihadwiki'.

[vii] Ibid.

[viii] For example, within a week's time, one single IS channel increased its membership from 5,000 to well over 10,000. Weimann, G. (2016). 'Terrorist Migration to the Dark Web' (2016) *10 Perspectives on Terrorism*. p. 42.

[ix] Clifford, B., and H. Powell (2019).George Washington Programme on Extremism, 'Encrypted Extremism: Inside the English-Speaking Islamic State Ecosystem on Telegram (2019).

[x] Macdonald S., and others (2019). A Study of Outlinks Contained in Tweets Mentioning 'Rumiyah,) 28 June. The Royal United Services Institute (RUSI).

[xi] The International Centre for the Study of Radicalisation and Political Violence (ICSR) (2017). Media Jihad: The Islamic State's Doctrine for Information Warfare.

[xii] Nagle, A. (2017). *Kill All Normies: Online Culture Wars from 4Chan and Tumblr to Trump and the Alt-Right* Zero Books.

[x] 111 Commission for Countering Extremism (2020). How Hateful Extremists Are Exploiting the Pandemic, p.7.

[xiv] For example, recent studies show that sarcasm alone can account for as much as a 50 percent drop in accuracy when automatically detecting sentiment: Sykora, M., Suzanne Elaya, and Thomas W. Jackson (2020). A qualitative analysis of sarcasm, irony and related #hashtags on Twitter. *Big Data & Society*, July–December, 1–15.

[xv] United Nations Office of Counter-Terrorism (UNOCT) (2021). Countering Terrorism Online with Artificial Intelligence. Available here.

Traditionally, in relation to the internet, CVE measures focus on combating the VE threat online, while PVE measures focus on preventative approaches to addressing the VE threat online: monitoring VE drivers online for early warning and action, building digital literacy, supporting the development of positive messaging online and promoting online dialogues aimed towards heightened societal cohesion.

The impacts of content moderation practices online are as follows:

- Content moderation is undertaken by social media platforms as per self-defined policy and practice. One major issue is that VE content can lack objective definition. However, creating definitions to work around does not necessarily solve the problem, because this can create boundaries for VE groups to manoeuvre. Far-right organizations in particular are often bound by in-jokes, memes and posts, which can be seemingly innocuous.[43] These groups make it purposely difficult to define content as illegal or harmful, and build their identity on the narrative that they, as early users of the internet, have 'deep knowledge' and are able to overwhelm efforts to censor.[44] Hate speech monitoring and community guidelines are purposely bypassed by inventing code terms and hashtags in order to avoid content being blocked.[45] As the Turing Institute found, "existing content detection technologies perform suboptimally at accounting for context, evaluating intent and recognizing irony, satire and humour [that] could have important implications for freedom of expression".[46]
- Additionally, 'harmful yet legal' content is often left online, which requires subjective removal, potentially threatening freedom of expression. Whereas illegal content is based around legal definitions that are applied through a lengthy judicial process that boasts the benefits of context and independent fact-finders, 'legal harmful' content does not have this established basis.[47]
- For many, social media is the only democratic arena in which many feel that their voice can be heard. By taking this away, individuals and groups may feel persecuted, which can push towards a disconnection from identification with the state, legitimize extremist claims against that state (or out-group) and encourage moves to smaller, harder-to-reach platforms and other forms of ideological expression. As noted by the Cyber Threats Research Centre (CYTREC), "Perceptions that regulation and removal targets particular social groups and identities, especially those who already feel silenced or marginalised, might exacerbate existing grievances."[48]
- National and regional bodies have increasingly developed online regulatory legislation to place legal restrictions on these practices. One major issue that remains is that legislation can be based on timeframes (removing content quickly), which can encourage overzealous removal to meet goals, place unrealistic pressure on small platforms that do not share the same resources for content moderation, and ignore more systemic issues.

How this relates to PVE practitioners:

- Content removal alone is not sufficient and poses a number of challenges to be addressed. Preventative measures are critical to address the VE threat online.
- A positive result of content removal as a whole is that major platforms become more difficult to utilize by VE groups, hence audiences for propaganda are lessened. However, VE groups/individuals become more difficult to pre-empt by law enforcement as the Dark Web and harder-to-reach platforms are increasingly utilized.[49]
- Content moderation practices and legislation can automatically remove related content available online for PVE practitioners to monitor and inform preventative initiatives.

Overstating the potential of online data

While the use of data for programming can be appealing due to the fact that it can potentially help better target programmes to prevent VE by unveiling trends online, the data available to PVE practitioners is limited due to necessary limits of personal data sharing by companies owning the data. Additionally, countries that experience challenges in internet access and/or in monitoring capacity will experience further limitations in reaching objectives through data collection and application.

43 See, for example, Evans, Robert (2019). The Boogaloo Movement Is Not What You Think'; and Tuters, Marc, and Sal Hagen (2019). M. Tuters, and Hagen, S. (2019). (((They))) Rule: Memetic Antagonism and Nebulous Othering On 4Chan. *New Media & Society*. Available here; Tuters, M. (2018). Larping & Liberal Tears. Irony, Belief and Idiocy in the Deep Vernacular Web: Online Actions and Offline Consequences in Europe and the US. Post-Digital Cultures of the Far Right. Colombia University; and Tuters, M. (n.d.). How Conspiracy Theories Spread Online – It's Not Just Down to Algorithms. Available here.

44 Bartlett, J. (2017). From Hope To Hate: How The Early Internet Fed The Far Right. *The Guardian*.

45 Ganesh, B. (2018). The Ungovernability of Digital Hate Culture. *Colombia/SIPA Journal of International Affairs*.

46 Vidgen, B., A. Harris, C. Dorobantu, and H. Margetts (n.d.)The Turing's Public Policy Programme Responds to the Online Harms White Paper.

47 Bickert, M. (2020). Online Content Regulation: Charting a Way Forward. Facebook, p. 9.

48 Patrick Bishop, P., and others (2019). Response to the Online Harms White Paper. *CYTREC/ Swansea University*, p.2.

49 For example, according to Jensen, et al. (2018), American extremists that post on social media have lower success rates than those that do not. Jensen, Michael, and Patrick James, Gary LaFree, Aaron Safer-Lichtenstein, and Elizabeth Yates (2018). The Use of Social Media by United States Extremists. START, College Park, Maryland. www.start.umd.edu/pubs/START_PIRUS_UseOfSocialMediaByUSExtremists_ResearchBrief_July2018.pdf

It follows that the countries with greater access to the internet and to platforms where open and free expression is permitted will be able to offer a greater volume of data. The operating framework is extremely relevant to the usefulness of data in a given context (i.e. is there censorship and are social media sites freely used by users or controlled by the state?). Certain languages and dialects are also underrepresented in monitoring tooling available. Therefore, an overreliance on online data would naturally act as a detriment to the countries challenged by the digital divide. Furthermore, access to the internet can be volatile due to both access-based and political factors, which can limit or disrupt the flow of online data for development purposes due to potentially unforeseen challenges such as internet shutdowns or power shortages. Accuracy of findings can be challenged by the fact that social media analysis may over-represent the views of particular groups within society who may utilize certain platforms more than others. The collection of relevant and comparable information can be challenged by the fact that the definition and identification of VE organizations can differ from state to state. Veracity of data is another consideration when determining the role data can realistically play in reaching objectives (i.e. the extent of data proxies, government interference and fake accounts).

Challenges faced by the use of online data in Sudan

Since Sudan's 2019 revolution, social media has been integral to the country's information landscape and culture of political expression. However, this has also opened up new opportunities for polarization, misinformation, conspiracy, a proliferation of hate speech, as well as for violent extremists to engage in online recruitment to their movements. Economically vulnerable women who do not typically have access to government services are especially susceptible to misinformation and recruitment by extremist groups. UNDP is using online data in order to understand the nature of this development and how it is manifesting online in order to assist audiences targeted in a more informed manner.

Working with local organizations, UNDP's Sudan country office is adapting Meekin, an award-winning AI platform developed by the social enterprise Koe Koe Tech, to the context and language of Sudan. The platform helps identify trends in social media around hate speech, and/or risk factors and early warning signs for extremism. The project will build on the platform's success in Myanmar, where the pilot project flagged 80,000 Facebook posts and comments.

The use of online data for delivering value is seen as particularly pertinent now as the pandemic has limited access to harder-to-reach communities by traditional means. The implementation phase of this project, however, has still faced unforeseen challenges such as a national energy crisis limiting internet access, national political upheaval leading to disrupted data collection, and reduced CSO capacity, making development of the dashboard's methodology and data collection hugely challenging. Hence, funding structures have had to be flexible with timescales. Given the challenge of accessing social media experts in the country with an understanding of context and language, capacity building has been a priority.

However, this project has ultimately placed UNDP as a thought leader in the country in the context of hate speech, which has, together with political developments causing the project to be of particular importance and relevance, received increasing attention. UNDP has therefore led in developing national capacities in this area through workshops and been able to build capacities based on its own learning experiences in undertaking this pilot project. The tool under development has demonstrated the ability to enrich contextual analysis and act as a solid basis to study emerging trends in social sentiment and influences, particularly for future political events such as elections.

CONSIDERATION 6:

Using both online and offline approaches

How can practitioners complement traditional data collection methods with online data to ensure that approaches mirror the online and offline nature of the journey to extremism?

The online/offline nexus of violent extremism

'Radicalization' can occur both online and offline, with one often reinforcing the other. Therefore, practitioners should use both new and traditional methods of data collection to understand the drivers of VE in a given context, and similarly should use data to inform both online and offline-based programmes to prevent the appeal of violent extremist communities and material online. A realistic understanding of the prominence of the internet in radicalization processes can guard against programming that is based on assumptions.

A focus on VE utilization of the internet should not, however, misplace the issue of 'radicalization' to the online sphere alone and perpetuate a false 'offline versus online' dichotomy.[50] Policymakers, practitioners and the media must take caution not to overstate the importance of the internet at the expense of offline factors. This can lead to an assumption of a false line of causation, perceiving the internet as a cause in itself, whereas it is more aptly described as a facilitative tool. A key finding from the UNDP-commissioned study, Social Media in Africa, is that while online influences play an important role in radicalizing individuals, "these have often been complemented by 'offline' physical influences in the form of in-person interactions with family, friends and other peer networks".[51] For instance, al-Shabaab, Boko Haram and Daesh's social media strategies centre around recruitment, but for many extremist group followers, 'offline' personal relationships also play a vital role. The importance of offline peer networks can be demonstrated by the fact that the geographical spread of violent extremists is not in line with internet coverage, but rather centres around geographic clusters, often around influential individuals, with the internet largely acting as an incubator for radical seeds.[52]

Therefore, it is important to know the trends found through online monitoring of extremist narratives, and hate speech trends leading to violence, as just one part of a larger picture that illustrates the threat landscape.

UNDP's offer in addressing online radicalization through P/CVE policies

How is UNDP addressing this through our programming? And what are the gaps in the programming? How can UNDP better address this problems identified?

Steps to develop wider 'online-based' PVE portfolios, e.g. counter/positive narratives, digital literacy, online engagement etc. – Competing systems of meaning created by UNDP:

UNDP has been building digital ecosystems through several initiatives such as developing guidance for online-based community engagement for PVE and building capacities to combat online and offline hate speech through partnerships, empowering youth, women and religious leaders to promote alternative narratives. In line with the call to develop "counter-narrative strategies" under Resolution 75/291, UNDP emphasizes the role of culture and faith in reinforcing the power of unity in diversity as the foundation of a cohesive, peaceful society.

50 Gill, P., and others (2017). Terrorist Use of the Internet by the Numbers. *Criminology & Public Policy* 16, p. 114; Whittaker, J. (2021). The Online Behaviors of Islamic State Terrorists in the United States. *Criminology & Public Policy*, 20.

51 Cox, K., W. Marcellino, J. Bellasio, A. Ward, K. Galai, S. Meranto, and G. Persi Paoli (2018). *Social Media in Africa*. Available here.

52 See, for example, The Soufan Centre (2015). Foreign Fights: An updated Assessment of the Flow of Foreign Fights into Syria and Iraq. New York. Available here. and Vidino, L., F. Marone, and E. Entenmann (2017). Fear Thy Neighbour: Radicalization and Jihadist Attacks in the West. Available here.

The false dichotomy of online and offline worlds can be particularly illustrated in young people, who are growing up in a world with unprecedented internet penetration

- ***Evidence of offline risk leading to heightened change of online risk in youth***

The importance of taking a longer-term, holistic approach to preventing violent extremism (PVE) can be observed in the fact that vulnerabilities of youth offline can increase their vulnerability to harmful content online. Evidence suggests that amassing social risk factors, such as poverty, family tensions, a sense of isolation, experience of racism or discrimination, exposure to violence and other negative events may increase a child's likelihood of involvement in an extremist organization.[i]Successful extremist organizations understand the power of capitalizing on lived experiences of inequalities, unemployment, exclusion and discrimination to build communities propagated by online narratives that utilize these experiences.[ii] This does not imply that risk offline leads necessarily to radicalization but children's behaviour in the digital world "is influenced by vulnerabilities already present offline and compounded by risks and harms they encounter online".[iii]

- ***How online hate can lead to offline attitudes***

UNDP Pakistan recently focused on attitudes towards women in online spaces, for example, by undertaking framework analysis of social media posts to understand the gendered online expression of certain current affairs-based case studies. The study found that as violence against women persists, it also mutates into new forms online and revitalizes the acceptability for gender tropes, policing transgressions and subverting the gains that access to democratized media platforms potentially promised women.[iv] The study points out the particularly harmful effect that this has on normalizing misogyny in young internet users, which is a known contributing factor to increased levels of violence in society, including VE, and points to anonymity offered by the internet and a dearth of adequate digital jurisprudence as two major contributing factors towards this problem.[v] The project used a data mining exercise to analyse the source and content of top Twitter trends targeting women during 'triggering' events in order to inform a targeted alternative messaging campaign to counter hypermasculine hate speech against women.

Many extremist groups use misogyny as an entry point for recruiting new members into broader hate groups. A study by UN Women in 2019 noted the high correlation between those who supported the right of men to commit violence against women, and support for VE more broadly.[vi] Where misogyny is supported and promoted by legislation and government policies, there is a notable link to the proliferation of hate speech against women, a rise in extremist groups and violence against women, both online and through physical violence.

[i] Littman, R. (2018). Children and Extremists: Insights From Social Science On Child Trajectories Into and Out of Non-State Armed Groups. United Nations University. Available here;
Bailey, P. (2015). Guidance for Working with Children and Young People Who Are Vulnerable to the Messages of Radicalisation and Extremism. Merton Safeguarding Children Board.

[ii] Umar, Mustapha (2015). Understanding the Complex Causes and Processes of Radicalization. Development. Development, Research and Projects Centre, Abuja: Office of the National Security Adviser of the Federal Republic of Nigeria.

[iii] Source: Katz, A. and Aiman El Asam (2019). Vulnerable Children in a Digital World. *Internet Matters*

[iv] UNDP Pakistan (2021). Masculinity and Online Abuse: Digital Discourses Targeting Women In Pakistan.

[v] ibid; Oslo Governance Centre (2021). Misogyny: The Extremist Gateway? Available here; and UN Women Asia and Pacific (2019). Misogyny and Violent Extremism. Implications for Preventing Violence Extremism. Available here.

[vi] Monash University and UN Women (2019). *Policy Brief: Misogyny and Violent Extremism - Implications for Preventing Violent Extremism.*

UNDP's programming to address online radicalization, hate speech, and information pollution incorporate both online and offline activities through three main trajectories:

Increasing understanding of hate speech and extremist narratives and their impact on societies through monitoring and analysis.

ONLINE

UNDP monitors hate speech and violent incidents as well as sentiment and social media analysis in order to understand grievances driving hate speech and motivating engagement with hateful extremist narratives. Monitoring is also used to better identify the instigators and outlets of hate speech.

Designing, identifying and scaling-up programmes to address the drivers and root causes of hate speech.

OFFLINE AND ONLINE

UNDP supports digital civic education to enhance media literacy as well as knowledge on rights and responsibilities online as well as critical thinking and the capacity for community engagement and intercultural dialogue in both offline and online spaces. Various initiatives, including Mental Health and Psychosocial Support (MHPSS), aimed to enhance community resilience from extremist narratives, from fake news to hate speech to the propaganda of violent groups, are also supported.

Offering alternative and positive narratives to counter hate speech.

ONLINE

Through the use of technology and the arts, and by leveraging the power of storytelling, UNDP supports creative individuals in developing and disseminating positive narratives of cohesive, peaceful societies aligned with local norms.

OFFLINE AND ONLINE

UNDP works with various civil society groups – youth, religious leaders, education actors, and media – as well as governments to not only raise awareness on issues of hate speech, but also to build the communities' capacities to promote peace, tolerance, and respect for diversity.

As outlined in the United Nations Department of Political and Peacebuilding Affairs' (DPPA) Checklist for Social Media Analysis,[53] when determining whether or not to undertake online and/or offline approaches, particularly in regard to conducting social media analysis to increase understanding of hate speech and extremist narratives and their impact on societies, the practitioner should first ask, "How relevant is social media in the country and region?" This would be followed by other questions, including:

- What social media channel is relevant, and why?
- Who is leading the debate on social media, how and why?
- What normative aspects need to be considered?
- Where do the news, media, tweets and comments originate from?
- What languages and dialects need to be considered?
- Is the post real or fabricated?
- What are key issues of debate (keywords), and major social media campaigns (usually hashtags)?
- How do online debates resemble or differ from debates taking place in the real world?

BuildUp, in its Social Media Analysis Toolkit for Mediators and Peacebuilders, offers three central potential uses for monitoring online data, which can help assess whether online methods are the most useful for the objectives of a programme.[54]

DAILY OR REGULAR MONITORING

Daily or regular monitoring is often the starting point to determine parameters for deeper narrative analysis or actor mapping. After deeper analysis, daily monitoring can be used to be updated on specific keywords (e.g. around a specific topic) or actors who are relevant to programming.

NARRATIVE ANALYSIS

Narrative analysis serves to understand the major trends in arguments around a topic or event, focusing on what is being spoken about and how.

ACTOR MAPPING

Actor mapping serves to understand the major influencers around a topic or event, focusing on who is connected to whom and what behaviours they use to exert influence.

53 UN-DPPA Innovation Cell – Social Media Checklist.pdf (reliefweb.int).
54 Main use cases for mediators and peacebuilders. Available here.

CONSIDERATION 7:
Utilizing online and AI tools

What tools are available to PVE practitioners and how can AI benefit projects that utilize online data for PVE?

In order to select a relevant tool, practitioners must decide which platforms are most relevant to monitor for the purposes of their objective, using relevant resources and literature,[55] potentially confirmed through interviews with active social media users includeinformed by the following questions:

- Who uses a platform? What are cultural norms for platform use? Do political elites use one platform more than others? Do women and men use platforms differently?
- What are use types (information sharing, opinion sharing, connection with friends/family, etc.) for each platform?
- What kinds of topics tend to be discussed on each platform? How relevant are they to the conflict context?[56]

Snapshot of online data collection, analysis, and visualization tooling available

Tools must be analysed in terms of whether the features they offer meet the objectives of the project. In order to reduce the burden on individual staff/teams, organizations often establish organization-wide agreements for accessing tools, which increases the likelihood that they meet the requirement. CrowdTangle,[57] for example, is a public insights tool, which is freely available to Facebook partners only and tracks influential public accounts and groups across Facebook, Instagram, and Reddit. This also includes all verified users, profiles, and accounts such as those of politicians, journalists, media and publishers, celebrities, sports teams, public figures. Organizations that may not be able to utilize free tooling can also create organization-wide subscriptions to tools such as Brandwatch.[58] This is a social media analytics tool that tracks data from a range of sources—such as blogs, news, forums, videos, reviews, images, and platforms (Twitter and Facebook) — decreasing the time cost of projects that may benefit from tooling. However, the financial resources required of such commercial tooling can act as a barrier to entry to this type of collection, analysis and reporting. Once a tool has been chosen, a list of search parameters (such as key words and/or actors) will be needed (usually by manually searching platforms chosen) to inform the data collection process. Parameters will be further condensed by geographical/time limitationson data available. These search terms can be used to create alerts on tools such as CrowdTangle or TweetDeck, and can be compared (the relation of keywords to key actors) and monitored over time to gain qualitative observations on how these terms are interacted with online.

Regardless of the tool or dashboard being used, social media analysis will require the following four-step process to be undertaken by the project's team:[59]

1. Decide which data to work with.
2. Collect that data from the social media platform(s).
3. Store and organize the data.
4. Look for patterns in the data, including by using visualizations.

While organization-wide agreements reduce the burden on practitioners, limitations persist when specific needs may be better met by other tools available, which practitioners do not have access to, pointing to the need for flexible tooling options. For example, in some contexts, certain platforms or languages monitored and analysis methods (e.g. sentiment analysis, indexing by key words, influence science, etc.) may be more useful than others. The usability of tools is also a

55 See, for example, Media Landscapes: Expert Analysis of the State of Media, medialandscapes.org
56 Main use cases for mediators and peacebuilders. Available here.
57 'Crowdtangle | Content Discovery and Social Monitoring Made Easy (2021). *CrowdTangle*, Available at www.crowdtangle.com.
58 Brandwatch (2021). A New Kind of Intelligence. Available at www.brandwatch.com/platform
59 Tick List from Build Up's *Social Media Analysis Toolkit*. Available here.

crucial consideration depending on the user of such a tool, for example, the difference between the required design for a data science expert utilizing a tool for specific research, compared to a practitioner looking to use a tool for day-to-day use. UNDP has made efforts to address the issue of finding tooling that meets the specific needs of its mandate while remaining accessible and readily useable for practitioners by developing a tool that can be adapted to the needs of the team's objectives within UNDP. The Crisis Risk Dashboard, for example, is a tool for data aggregation and visualization to support contextual risk analysis conducted by UNDP and the wider United Nations system.[60] This tool can use already generated data and lexicons (or the developing team can support others within the system to develop these) and tailor them to a project's objectives.

External tools

a. Example of a free public insights tool: CrowdTangle

CrowdTangle* is a public insights tool from Facebook, which tracks influential public accounts and groups across Facebook, Instagram, and Reddit, including all verified users, profiles and accounts such as those of politicians, journalists, media and publishers, celebrities, sports teams, public figures. CrowdTangle does not track any private accounts or posts made in private groups. Its full features can be utilized by those who qualify as partners of Facebook. CrowdTangle also can track seven days of public Twitter data via CrowdTangle Search and the Chrome Extension.

Organizations primarily use CrowdTangle to:

Follow. Follow public content across Facebook, Instagram and Reddit.
Analyse. Benchmark and compare performance of public accounts over time.
Report. Track referrals and find larger trends to understand how public content spreads on social media.

Available features

- Is free.
- Has excellent developer support.
- Intuitive Application Programming Interface (API) (connection already built).
- Has pre-built metrics on user interaction/trending posts.
- Able to pre-build lists of groups and pages to monitor.
- Has an excellent Search Function.
- Allows to search image text (e.g. analyse memes).

Restrictions

- Does not include any data on Facebook comments.
- Only allows to monitor public groups and pages.
- There have been reports of Facebook executives requesting to filter Crowdtangle's data to remove any hateful or negative content from the platform. This raises questions regarding the reliability of the data.
- Some API features require an advanced request form and there are API rate limits.
- Built-in analytical functions are for digital marketing purposes and are not always applicable to the typical use case.

Table Content from Social Media Analysis Report, UNDP Crisis Bureau Crisis and Fragility Policy and Engagement Team (CFPET), Available here.

* Crowdtangle | Content Discovery and Social Monitoring Made Easy (2021). Available at www.crowdtangle.com

60 UNDP (2021). Crisis Risk Dashboard. Integrated Solutions for Sustainable Development. Available at sdgintegration.undp.org/crisis-risk-dashboard

b. Commercial social media analytic tool example: Brandwatch

Brandwatch is a social media analytic tool that tracks data from a range of sources such as blogs, news, forums, videos, reviews, images, and platforms (Twitter and Facebook) and allow brands and companies to understand consumer insights, trends, influencers, and brand perception.

Data collection

Crawl: Brandwatch obtains content using proprietary web crawler technology that crawls a list of sites that it covers.

Analysis

These pages are then analysed by techniques such as indexing, sentiment analysis and influence science.

Indexing: The search index technology puts the contents in an index so that clients can search them by words.
Sentiment analysis: Brandwatch uses natural language processing technology to conduct sentiment analysis. It is a mathematical model that models linguistic features that indicate sentiment. In addition, there is a rule-based process to help the software better understand how context can affect sentiments. As language evolves, the dictionary that machines use to comprehend sentiment will continue to expand.
Influence science: The Brandwatch Influence Score measures an individual's ability to generate engagement and amplify messages.

Available features	Restrictions
• Multiple curated sources of data (at least Facebook and Twitter). • Sophisticated analysis layers that include insights on keywords/trends/topic modelling/dialogue drivers, etc.	• Relatively high price. • Brandwatch is designed for marketing and monitoring brand reputation. Therefore, it is not a seamless fit for the usual social listening applications (e.g. monitoring hate speech).

Table content from Social Media Analysis Report, UNDP Crisis Bureau CFPET.

c. Free social media analysis tool example: Phoenix

Build Up has also developed an open-source, non-commercial social media analysis tool designed to meet the needs of peacebuilders and mediators who want to work ethically with social media. Currently, Phoenix is using public Facebook posts from pages and groups, together with manual assisted scraped comments, Tweets from handle list of keyword search, YouTube videos from channel list or keyword search and comments from channel list, together with events from ACLED.

Sources:

Brandwatch (2021). Brandwatch: A New Kind of Intelligence. Available at www.brandwatch.com; UNDP Crisis Risk and Early Warning TeamSparkblue (2021). Available at crowdtangle.com ; Phoenix demo: UNDP CFPET (November 2021). Social Media Monitoring and Analysis – Tools and On-Going Initiatives. Available on YouTube.

United Nations tools
Examples of in-house tool development

a. The Crisis Risk Dashboard

The Crisis Risk Dashboard (CRD) is an internal platform developed by the Crisis and Fragility Policy and Engagement Team (CFPET) (part of the UNDP's Crisis Bureau), which helps monitor and analyse contextual risks for UNDP and UN system decision-making. The objective of the CRD is to help UNDP prevent and respond to crisis risk at the global, regional and country levels by providing an evidence base on contextual risk through relevant and timely data. In the context of PVE, the CRD provides a customised platform that integrates indicators and datasets to monitor how a country's risk profile evolves over time.

Relatedly, CFPET has developed several solutions to social media monitoring / social listening that are integrated within the Crisis Risk Dashboard. In responding to requests from United Nations Country Offices and Resident Coordinator Offices, the team has co-designed and built standardised analyses, drawing inspiration from available tools. These include several sentiment analysis and hate-speech detection frameworks, as well as the capacity to build custom solutions and monitor specific trends, content, and influencers online. Most importantly, the CRD provides a platform to integrate these social listening tools with other datasets and analyses, further facilitating comparisons of indicators and improved contextual analysis for PVE.

In Sri Lanka, for example, under initiatives that aim to establish an early warning system, UNDP has worked to improve the knowledge base on countering hate speech and intolerance. The collected data is now informing critical assessments and risk management for programming and policy of the United Nations Resident Coordinators' Offices and UNDP through the UNDP Crisis Bureau.

b. Qatalog

Qatalog[i] developed by United Nations Global Pulse for members of the United Nations system, is an AI-powered tool that offers features for data mining social media posts, including the ability to pull and analyse Facebook data using manual annotation techniques in addition to automatic geolocation, translation, and machine learning-driven text classification. Tags can be then automatically applied to posts and data can be visualised and collected into microsites for analysis. Raw data can be downloaded for further analysis.

c. Sparrow

Sparrow[ii] is a user-friendly driven tool developed by United Nations Innovation to generate a simple outlook into Twitter trends in a particular country. The data that is gathered is curated based on the accounts that one Twitter user is following, and so a project team must create a Twitter account specifically to follow targeted accounts for analysis.

Notes:

i UN Global Pulse (2021). Qatalog – An Analysis Tool for Insights into the SDGs. Available here.

ii Mysparrowreport.org (2021). Report Dashboard. Available at mysparrowreport.org/about.

The UNDP Digital Lighthouse initiative (Amman Hub) invests in research initiatives that act as a 'lighthouse' for similar programmes across the organization. One project is developing machine learning tools, and leveraging AI to identify, classify and monitor hate speech in the Arab Twittersphere, using Tunisia as the initial pilot country in the Arab region.

This project aimed to:

- monitor and analyse hate speech trends on social media, specifically, Twitter, in the Arab region;
- leverage knowledge from the social sciences on the antecedents, root causes, and predictors of hate speech, and detect and predict these precursors in instances of hate speech on Twitter;
- engage with social media (Twitter, specifically) as a platform to monitor and understand the scope of hate speech in the Arab region;
- leverage innovative technological advances, such as AI and machine learning, to detect and monitor hate speech content in the Arab Twittersphere.

The project methodology consisted of five distinct phases: (i) preparation and set-up (team set-up, conceptualization and literature review, infrastructure setup and design of data pipeline, tweet mining, and annotation schema); (ii) setting up the Semi-Automatic Interactive Classification and Annotation (SAICA) platform; (iii) pilot testing of the coding scheme; (iv) Tweet annotation and classification; and (v) data analysis and validation, reporting, and visualization.

The project has faced major challenges in meeting its objectives due to the resources needed to carry out a project involving big data and AI, which aimed to develop a tool from this process. In addition, the data made available by Twitter made it difficult to reach the stated objectives of the project. The project had to therefore adapt the codebook used to label data from monitoring hate speech to 'negative speech'. Despite this change in methodology, the objectives of the project were still met, and much useful data was found in this area of speech that fell just outside of Twitter's content detected as hate speech and therefore removed from the platform before it could be used by this project.

Developing tools through transparent processes

Automated tools that lack principled processes can create heightened chance of inaccuracy in results, which for PVE can lead to misinformed programming. Ultimately, AI only scales up the research it is based on; hence, if a tool lacks rigorous processes of development and verification, the tool will be unreliable at best and at worse, harmful. As AI solutions in this space are heavily dependent on the underlying data that the model is trained on, there is a greater need for rigorous content definitions and annotation methodology. However, perfect data does not exist, hence bias will always be present. Fully removing the bias should not be the aim, but rather, identifying it and minimizing it is the most optimal route.

Using AI and big data in a principled manner

AI tools can be used to enhance the volume of data collected, tailored to reaching the ultimate objectives of the intervention.

Artificial intelligence

Artificial intelligence (AI) is the simulation of human intelligence processes by machines, especially computer systems, which can be used to process larger quantities of data (also referred to as 'big data') and discover patterns in data unseen to the human eye. Machine learning and deep learning are examples of subsets of AI. An AI system may, for example, include sensors that capture the environment plus a machine learning algorithm to perform a certain task based on the data received. Machine learning concerns algorithms that can learn from data, i.e. progressively improve performance on a specific task.[59] In this context, AI can refer to certain tools utilized by PVE practitioners and partners in order to organize and amplify online data collection processes. Computers must be taught by humans via a chosen methodology that is coded into computers in order to artificially produce the chosen sequence of data collection and analysis from those sources (e.g. social media platforms) deemed most relevant.

61 United Nations Office of Counter-Terrorism (UNOCT) (2021). Countering Terrorism Online with Artificial Intelligence. Available here.

In order to utilize AI in a risk-informed manner, an understanding of the potential of AI and the need to consistently monitor the process when building and using tools is essential.

Though there are differences between approaches to monitoring content shared in violent extremist spaces online and approaches to moderating and removing terrorist content, including content that promotes terrorist acts, AI has been increasingly utilized by both PVE and CT actors. While the following recommendations outlined by UNCRI and UNCCT are provided for law enforcement and counter-terrorism agencies in South Asia and South-East Asia, PVE practitioners should note the lessons learned:

Acknowledge the limitations of AI:

- AI is getting more effective and accurate with enhanced understanding by humans of how to best develop it but is not perfect and may never be. There still exist many technical challenges.
- Humans cannot fully grasp all contexts; expecting more nuanced understanding from AI would therefore be ill-informed.
- There is no 'one size fits all' solution. AI tools must be tailored to specific contexts and requirements.
- AI presents results as probabilistic calculations and not infallible predictions.
- AI can make predictions; it cannot give meaning to the results, and unforeseen connections (unintended neural connections) in highlighted patterns can create errors in results.
- AI cannot unilaterally eliminate terrorism and VE online. Investments into AI should be flanked by efforts to prevent radicalization and VE at its roots.

Ensure human oversight and accountability

- Thorough consideration should be given to how AI-enabled technology can be used to augment human analysts as opposed to replacing them or their functions.
- Automated decision-making should not be encouraged. Analysis and conclusions reached on the basis of AI systems should always be reviewed by a trained human, who should make the final determination.
- All personnel working with AI-based systems should understand the system, and training should be provided to personnel on technical and legal aspects of the use of the tools, as well as possible associated human rights risks.
- Oversight mechanisms should be established and equipped with sufficient technical expertise to be able to understand and appropriately assess the use of AI. These oversight mechanisms should regularly assess each instance of the use of AI systems, the legality, proportionality and necessity of their use, and the quality of the outcomes they produce. Institutional checks and balances should be guided by transparent policies.

Source:
Excerpt of two recommendations from Countering Terrorism Online with Artificial Intelligence, a joint report by United Nations Interregional Crime and Justice Research Institute and United Nations Counter-Terrorism Centre.

The potential of AI tools

Principles of tool development

AI tools can be used to enhance the volume of data collected, organized in a manner that lends itself to the ultimate objectives of the intervention. Fairness, transparency, proportionality, accuracy and accountability are central principles when developing useful and responsible ethical AI tools for scaling up and/or streamlining data collection and analysis processes. However, these ethical principles must be met with transparent M&E mechanisms to ensure they are practically and meaningfully enacted into project lifecycles. For example, to be classed as 'fair', it must be demonstrated that algorithmic decisions do not create a discriminatory or unjust impact on the end-users. While the technology behind AI tooling will entail a neutral statistical and mathematical process, AI can amplify biases of societies or within the teams developing the tool when trained with (intentionally or unintentionally) biased datasets. This may result in incidents of automated solutions discriminating against individuals, groups, and/or communities on prohibited grounds.[62] Linked to this is the need for transparency and explainability of algorithmic decisions and actions taken on the basis of such results, which can be aided by the use of explainability tools to ensure that end-users are able to interpret what elements used in the machine learning model were responsible for each

62 ibid.

specific outcome.[63] Limitations of accurate and ethical data collection and amplification can also stem from the extent and nature of available data, including potential gender, ethnicity, class, and/or demographic-based misrepresentation. While, generally speaking, simple models with larger data sets are more effective than complex models with small data sets, a larger quantity of data will not be useful if the quality is poor due to inaccuracy or due to data sets being unrepresentative or outdated.[64] A lack of available data on particular VE groups, or the inability of tooling to pick up the language of these groups may also create disproportionate focus on particular at-risk groups in a given context. To overcome this, regulation and monitoring of access and use of AI tooling is needed.[65] For specific interventions, diverse and multi-sectoral teams with contextual knowledge and technical expertise in design can be consulted to enhance the accuracy of this process.

Low-resource languages, the nature of online dialogue and how quickly it evolves over time are key challenges to be solved when deploying AI-based solutions in this space.

Snapshot on how AI can both offer opportunities and pose changes: Gender and violent extremism

Investing in online-based research projects is crucial because the internet has created a new space for violence against women to take new forms, reducing advancements in society and normalising toxic gender roles and rhetoric, particularly in youth.[66] Toxic gender roles can be heightened by online rhetoric and have been seen as an important element in the building of identity and community among many violent extremist groups.[67] Understanding the contextualized reality here is crucial to designing projects and policy to mitigate the real harm of these advancements, since technology companies and national/regional governments have been unable to effectively regulate online activity to eliminate harm to women. In addition, anonymity online, which is a contributing factor in harming women in this context, is unlikely or impossible to be banned. However, particularly when undertaking social media monitoring using big data and AI tools, gender-based misrepresentation can be a risk due to potential factors, such as limited access to the internet in certain contexts, and therefore a lack of data collected from women users, or a lack of gender diversity in the project monitoring teams that thus may have weaker capacity to identify related issues of bias. AI will only amplify the research that it is based on, hence data and processes of data collection that misrepresent women will only be amplified through AI tooling. This explains why it is important to have diverse teams with local knowledge (including language/dialect) ideally involved at all stages of project development. When undertaken with rigorous and responsible processes, AI tools can help provide live insights into gender and VE, communicated in a way that is easily accessible and relatable to a team's/project's objectives. However, due to limitations of access, which can be amplified for women in many development contexts, online data-driven projects should not be depended on, but rather should build on (together with other emerging data collection-enhancing methods such as behavioural science) traditional methods of data collection and knowledge building to ensure that women's experiences are accurately portrayed in research and subsequent programming and policy.

63 Kaur, H., and others (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. CHI 2020 Paper. Available here; United Nations Office of Counter-Terrorism (UNOCT) (2021). Countering Terrorism Online with Artificial Intelligence. Available here.
64 ibid.
65 See, for example, OCHA's peer review framework for Predictive Analytics in Humanitarian Response: Available here.
66 UNDP Pakistan (2020). *Masculinity and Online Abuse: Digital Discourses Targeting Women in Pakistan*. Available here.
67 Pearson, E. (2018). Online as the New Frontline: Affect, Gender, and ISIS-Take-Down on Social Media. *Studies in Conflict & Terrorism*, 41(11), 850–874. Available here.

CONSIDERATION 8:

Applying findings to create impact

How can practitioners design projects with target beneficiaries in mind and communicate findings in a way that leads to desired action by appropriate stakeholders?

In order to ensure that knowledge gathered through the data cycle creates a maximum desired impact through programmes and policy, clear and relevant communication to stakeholders must be thoughtfully considered. Data can seem difficult to decipher and intangible unless directly and clearly related to the needs of potential users/readers. Continuous reporting of data in pre-determined timescales can help end-users keep updated with findings and follow a narrative of learning if communicated clearly and coherently. While explaining methodology is important, an over-reliance on the technicalities of the data cycle can separate users from the impact of data. Therefore, a combination of clear description of methods used, risks assessed and the human impact of findings, can help create a holistic and approachable overview of a project. A communication plan should include how the findings will be communicated, to which stakeholders, and according to which objectives.

A key part of communicating findings is through visualization of the data analysed. Although visualizing data can communicate and break down complex analysis in a more palatable format, such visualizations need to be high in quality, low in complexity, and used where only clearly useful to demonstrate a point and avoid alienating audiences. Suggesting how these findings can apply to stakeholders to meet both short term and long-term policy objectives can help encourage findings to be taken up and recommendations acted upon. Communication methods of findings will differ with the objective and subsequent analysis method used. For example, bar or pie charts will best demonstrate quantitative analysis, while network maps will most accurately communicate network analysis.[68] AI can be advantageous to creating reporting and dashboards that blend data visualizations with meaningful narratives in order to clearly communicate data in a way that humanizes the statistics at hand. These information products can be built on the needs of the user.

68 See Build Up's Social Media Analysis Toolkit for guidance on interpreting patterns found: available here.

Example of creating projects with consideration of impact from the design stage

UNDP Bangladesh devised a project that seeks to build local research capacity, integrate the findings into project activities and input lessons learned into national, regional and global efforts towards preventing violent extremism (PVE). Data obtained through this project has increased the reach and impact of pro-peace and tolerance narratives in target audiences, reduced engagement with negative violent extremism (VE) content, and has increased multi-stakeholder national understanding of VE trends.

- The project's stated objectives were to understand the Bangladeshi audiences who are attracted to extremist narratives and whether economic inequality, development, or human rights concerns in Bangladesh or among the Bangla-speaking diaspora shape violent or exclusionary narratives online. A *risk assessment* was carried out together with their partner, SecDev, and was based on taking a public health and expert-driven social science approach.
- *Analysis* is organized and communicated into regular reports, produced in a way that is most relevant for selected audiences. A dashboard is being developed in order to make data collected more readily accessible and automatically organized in line with project aims.
- The live insights gained provide ongoing programming support for both online and offline PVE initiatives by *applying* understanding acquired of the range of violent extremist and exclusionary rhetoric in Bangladeshi cyberspace, outlining which issues are most salient to online communities of Bangla-speaking sympathisers.
- UNDP Bangladesh's partner SecDev is producing monthly and quarterly reports with monitoring findings and analysis. Internally, these findings have been used to orient youth partners on strategies that VE actors are following in order to inform their responses. UNDP has also used SecDev findings to formulate problem statements for IdeaLabs and hackathons (called Digital Khichuri Challenges). Both government and non-government actors in PVE are receiving the SecDev briefs and 3–4 times a year, presentations and discussions are organized with PVE actors on the findings.

The impact of this work can be seen in the following outcomes:

- Increased reach and impact of pro-peace and tolerance digital narratives: Digital narratives activities supported by the project were fine-tuned on the basis of SMM data, reaching more than 18 million people directly, including more than 10 percent of Bangladeshi Facebook users in the targeted group: young men between the age of 18 and 24 identified as most at risk of identifying with VE narratives. Furthermore, these alternative narratives had the greatest engagement with this target group.
- Positive response from government, civil society and United Nations entities, which also use this data to inform their policies and investments: For example, UNICEF, the Australian government and Bangladeshi government officials affirmed the role of this data in informing their own work. Several major media outlets including The Financial Express, Daily Star, Prothom Alo, Jungator, Jaijaidin, New Age and Banglanews have featured this data, including as part of newspaper headlines. A survey showed that 40 percent of Bangladeshi government, civil society and international community actors regularly receive and read Partnership for a Tolerant, Inclusive Bangladesh (PTIB) data products.
- High level of engagement with this content: The opening rate of the project's reports is several times the industry average.

@UNDP Pakistan

MONITORING, EVALUATION AND LEARNING

CONSIDERATION 9:

Creating M&E processes for data collection and application

How can practitioners create more effective M&E processes with online data to better inform future projects?

There is a need for M&E frameworks to be tailored and updated to incorporate online data from these new sources, including online platforms, as part of the evidence base to monitor progress and evaluate the impacts of PVE projects. Furthermore, the more rigorous measurement approaches common to behavioural insights may also be used to measure the impact of initiatives that take place predominantly in digital spaces. For example, different evaluation methodologies can be utilized to measure the impact of digital citizenship initiatives on resilience to VE. Indicators based on monitoring online/social media content can also be included in the M&E plan of projects to support the ability to identify and effectively respond to hate speech trends observed online. A-B testing or pre-/post-surveys may be used to improve the effectiveness of online alternative narratives campaigns. These various M&E processes should be put in place to ensure that objectives are met and methodologies revised where deemed ineffective through iterative and adaptive programming.

Defining data collection needs: The data that a programme collects is determined by its objectives (linked to its Theory of Change). While global policies and frameworks identify general 'push and pull' factors believed to be conducive to VE, it is important that UNDP programming be contextualized to localized drivers. These can be identified by research, conducted through both traditional offline and online methods. For example, local context, subject matter, and language experts (including local CSOs active in PVE) can use their knowledge to frame research questions, conduct and/or participate in data collection and analysis on local drivers of VE. Their expertise can frame data collection in online spaces; leveraging AI to gather data from potentially trillions of data points to assess the prevalence of these issues, the demographic characteristics of the populations most affected, and their geographic locations. For example, local context, language, and subject matter experts may identify key themes (e.g. grievances) exploited by violent extremists, as the basis for digital research that frames keyword searches to identify how significantly these narratives resonate in the country, and which populations are most attracted to them. AI-driven data collection can identify the overall numbers, as well as the specific demographic populations, most affected by those grievances: i.e. the sex, age and location of people who search for, engage with, or produce content that expresses these grievances. Alternatively, context and subject matter experts can name VE actors that enjoy support in the country or region in which programming is taking place. AI-driven data collection can then assess the reach of these actors, extent of engagement with their content, and to some extent, the impact of this content on attitudes and behaviour in online spaces. The data on issues to address and populations affected by these issues form the core elements of a quality M&E system:

1. **The problem:** What programming aims to achieve, or change(s) it should seek to make (i.e. project objectives).
2. **The population:** To whom programming should be targeted (age, location, sex of persons attracted to violent extremist narratives (i.e. project beneficiaries).
3. **The present situation:** Specific information on the extent of the problem, which can be used to frame project baselines and targets), for example, data on the numbers of people who actively seek violent extremist content; proportion of the population who express grievances that are exploited by violent extremist actors or are known drivers of radicalization.

Analysing data: Core information on what issues to address through programming, and whom to target programme activities towards provides the starting point for further enquiry on how to effectively reach, generate attention of, and even attempt to change the attitudes and behaviour of the target audience. When the core at-risk population or beneficiary population has been identified (from an analysis that draws on data from across an entire region or national geography), it is possible to analyse this data and generate more insights to shape programming approaches. For example, information on the types of content that target users look for (e.g. combination of words, sounds, images; entertainment, sports, and political analysis), messengers and messages that they engage with most, and where they look for them (e.g. platforms, channels, blogs, and forums) can be used to frame alternative narrative content development and delivery. It is also possible (albeit to a limited degree due to cost) to measure to what extent attitudes change as a result of exposure to specific types of content by identifying whether and how sentiment expressed in digital spaces changes following exposure to that content (and, again, which types of content have the most significant effect on attitude). With this information, UNDP can design evidence-based activities that respond to the preferences of the specific audience vulnerable to VE.

The AI-enabled digital data collection, grounded in the 'offline' experiences of local experts, thus informs targeted programme design: 'what' issues to address; 'whom' to target; and 'how' to effectively engage and influence them. By repeating these data collection exercises before and at given periods throughout programming, UNDP can identify whether or not, for example, overall volumes of searches for violent extremist content changes; whether engagement with violent extremist content is reduced among the targeted population; and whether attitude towards a given grievance (e.g. migration, corruption) improves among the targeted population. UNDP can also identify which types of content has the biggest effect and refine its approach around these contextualized good practices. See, for example, the text box on UNDP Bangladesh's social media monitoring and alternative narratives initiative above. As part of UNDP's programming policy support function, this data can also inform offline programming and policy development to respond to vulnerable populations and grievances identified through data sourced through AI. See also, for example, the UNDP Sudan example above on how AI and online data are used to identify and empower economically vulnerable women.

Enhancing traditional data collection methods through measurement: The potential of online ecosystem mapping

Supported by Facebook and the European Union, in Southeast Asia, the regional programme launched the #ExtremeLives website as a storytelling platform, using a variety of communications-focused or storytelling-focused media to increase the awareness of the factors that drive extremism in Asia among youth audiences. To date, Extreme Lives content has been developed based on findings of research identifying drivers of extremist violence in the region, and consultations with experts (e.g. UNDP Country Office staff and civil society organizations working on the ground). This data has been used to identify topics and speakers for the videos (i.e. by providing clarity on who is vulnerable to radicalization, and suggesting which topics they may be interested to learn about, and what types of content may address the beliefs, attitudes and behaviours that increase their vulnerability to violent extremist content). After identifying the audience and issues through this 'research review and consultation', the relevance of proposed content was validated and strengthened by a documentary company, which conducted further desk research to ensure that videos respond to local contexts. .

The advantage of this approach is that it promotes ownership and engagement of CSOs in identifying themes and target audiences. CSOs consulted may propose themes that they are working on and familiar with. However, in some places research reports are few in number, contain information that is rapidly out-dated, and almost always draws on a smaller quantity of data than that available through social media monitoring. Therefore, UNDP cannot be sure that the right topics, audience and delivery channels have been identified. In addition, the existing measurement of impact (e.g. number of views, likes, shares, minutes watched) does not indicate whether and how Extreme Lives content is

changing attitudes and behaviour of those people it does reach; nor is it possible to draw a credible link between this type of data and national-level indices on levels of tolerance, attitudes to diversity, trust etc.

To address these issues going forward, UNDP is strengthening the Extreme Lives video campaign through investment in measurement activities to help ensure that content in its next series: (i) is targeted and disseminated to its intended audience of people vulnerable to violent extremist narratives; (ii) achieves a measurable impact on the attitudes and behaviour of this 'at risk' audience; and (iii) is regularly fine-tuned based on data to continually enhance targeting and impact.

The Extreme Lives measurement strategy is undertaken through the following:

1. **Digital ecosystem mapping** to identify *who* is looking for VE-related content, *what* grievances they have, and *where* they look for it. With this information, the content messages, messengers and mediums are targeted
2. **Embedding behavioural prompts** (e.g. access to psycho-social support or youth engagement services) and tracking how many viewers respond to this 'call to action.'
3. **Facebook/Google A-B testing** to fine-tune content in response to feedback.
4. **Focus group discussions and pre-post surveys** with participants to measure whether and how their attitudes change as a result of exposure to Extreme Lives content. Participants are followed up after six months to see whether and how behaviour has changed, and attitude changes are sustained.
5. **'Gamified' online surveys** to measure whether and how viewers' attitudes change after having seen Extreme Lives videos.
6. **Digital measurement activities** to assess engagement with content, to what extent it reaches the targeted populations identified through the digital eco-system mapping; whether it reaches VE-dominated digital spaces, and whether it changes attitudes to diversity in these spaces.

All digital monitoring activities conducted as part of this initiative are being conducted by GDPR-compliant companies, and in consultation with local civil society organizations. This consultative element aims to ensure that local actors are involved in analysis of the data and can use it to inform their own advocacy and action. Furthermore, the companies carrying out this work are required by UNDP to conduct workshops to ensure knowledge transfer to civil society organizations on social media monitoring expertise, and to avoid and address any biases in algorithms. Extreme Lives forms part of a regional programme that also includes in-country programming to address VE (Philippines, Malaysia, Thailand, Indonesia). These programmes use a variety of methods, including data from resilience surveys, community-based early warning systems, and outcome harvesting, to validate and complement data from online and AI sources. UNDP deems this critical to address any bias or inherent limitations in the data derived from online sources.

CONSIDERATION 10:

Ensuring that learning feeds into future programming and policy

How can we learn from the data to inform our future PVE programmes and policies?

For learning from the use of online data and AI to influence future programming and policy, it is necessary to establish processes to analyse, communicate and share this data with all who can benefit from it. Typically, those who can benefit include actors across the 'whole of society' (civil society, government, academia, business, United Nations agencies, international NGOs and other international development partners). Entities from each sector typically have different capacities to contribute to prevention of VE, and the data that comes from online monitoring efforts can help them to leverage these respective capacities for this purpose. For example, as described above (text box, Consideration 8), data from monitoring efforts supported by UNDP Partnerships for Tolerant, Inclusive Bangladesh (PTIB) project has been used to inform the design of CSO counter-narratives; United Nations humanitarian response and government allocation of resources to respond to the Rohingya crisis, quality media reporting during COVID-19, international development partners' strategy to invest in Bangladesh, as well as foreign direct investment of global business partners.

For data to be used for this purpose, key consideration must be given to:

Engage partners in analysis: It is necessary to engage diverse partners in the analysis or 'sense-making' processes that make data meaningful to, and usable by their intended audience. In addition to providing essential information (from their own communities, institutions, and engagement) that can contextualize the data and draw relevant meaning, involving partners in analysis supports: (i) the identification of practical actionable recommendations that help to ensure that data informs policy, programming, and investment decisions of diverse institutions; (ii) awareness of the monitoring process as a useful input to such decisions; and (iii) ownership of the monitoring processes, which serves to encourage use of relevant data by those institutions.

Invest in quality communication: Data does not 'speak for itself': often, it is easier to understand the real-life significance of issues by sharing key statistics (e.g. on macro-level issues such as the level of misogyny in a country, or the prevalence of grievances against particular groups) with a single story or 'micro-narrative' from one person affected by the issue. In addition, according to UNDP experience, the most effective way to ensure that data is used by its intended audience is: (i) to produce short reports (of 1–2 pages only) that are easy for time-starved audiences to consume; (ii) use infographics that present key data 'at a glance'; and (iii) share data frequently (e.g. bi-weekly reports or bi-weekly short reports), which makes it easier to create a habit of using data as part of daily work.

Distribution: Data has the capacity to unite diverse actors around a common evidence-base for action. UNDP is ideally positioned to leverage data for this purpose, due to its status as a trusted partner of government, as well as civil society, private actors and other UN agencies. UNDP can use this network to distribute relevant data to all actors with mandates and capacities to contribute to prevention efforts as well as utilize the data for advocacy for policy reforms. In addition, UNDP is a lead provider of support to the development of PVE National Action Plans (NAPs). To support NAP implementation,

governments are increasingly requesting UNDP's assistance to develop PVE NAP M&E frameworks, with participation of actors from across the 'whole of society' to provide necessary data. These frameworks are supported by government-administered coordination units, which collect, analyse and use data to guide PVE NAP implementation. These government-led PVE NAP coordination structures are ideally suited to distribute online data among all relevant entities. As national authorities, it is also possible for PVE NAP M&E framework coordinators to encourage (or require) entities to demonstrate how they are using such data to improve the targeting of their efforts to support NAP implementation (e.g. in terms of selection of how geographic areas, demographics, or themes that their activities address correspond to the priorities suggested by the data).

Conclusion

Collecting and applying online data to PVE programming is one approach to better understand the contextual drivers of VE and can complement other online and offline approaches towards creating systemic resilience against VE through strengthening the evidence-base for action. Utilizing online-based approaches for PVE should be understood as one of many tools to practitioners and policymakers and not as a silver bullet. To navigate the constantly evolving policy and technological landscape, there are key components which must be considered when developing technology for online data monitoring and analysis: (i) increasing understanding of the opportunities, challenges and risks; (ii) developing research and utilizing the findings to inform programming within both the online and offline spheres; and (iii) where possible merging both online and offline approaches, in order to take holistic, all-of-society approaches to PVE.

Approaches to counter VE and hate speech in response to the growing online VE threat must be complemented with preventative efforts, and in order to undertake this type of programming, an understanding of the multi-layered policy landscape is crucial. This includes due diligence considerations for partnership with the large technology companies, which have far more power over society than smaller private partners or a development-based organization. Irrespective of working with partners, or developing in-house methodologies and tooling, transparency is crucial at all steps of a project's development. This is to ensure not only that data collection is based on sound research, but also that methodologies are ethically sound and human rights-compliant, which is particularly relevant when the amplification effects of AI are being utilized. This can be analysed through tailored risk-assessment frameworks and guidance that take into consideration each step of the data process cycle (including actors involved in application of data), together with due diligence for partnerships. Similarly, findings must be carefully and strategically shared with necessary stakeholders with conflict sensitivity to mitigate risks and create impact.

Tailored frameworks, strategy and guidance based on established risk management frameworks and data principles for development can help practitioners build their capacity to make practical, ethical and human-rights informed decisions on methodologies, tools and partners for utilizing online data. Given the need for partners with resource capacity and technical expertise on online data, the international community must advocate for human rights compliance when forming partnerships and identify areas where their partners can be supported to reach these standards. Involving as diverse a range of actors within a community at hand as possible can alleviate many challenges arising from lack of local-level understanding of context and language. By building the capabilities of local actors, including peacebuilders, to navigate the online space and monitor VE narratives and hate speech using new technologies such as AI, UNDP can both empower communities and encourage sustainable practices for a human-centred preventative approach to VE. CSO partnerships should be prioritized while holding consistent consideration of the wellbeing of individuals working on projects that utilize potentially sensitive online data, a consideration that should be included in risk-assessment measures. Ultimately, effective risk management is needed to fully realize the potential of online data collection, analysis and application for PVE. While many challenges still exist, the experiences of the UNDP country office pilots illustrate that data has the capacity to unite diverse actors around an evidence-base for action. With the online space playing a particular role in the radicalization process to VE, risk-informed approaches to using online data can enable PVE practitioners to better understand and respond to the evolving online threat landscape.

@UNDP Sudan/Afra Tagelsir

ANNEXES

ANNEX 1:
Relevant Human Rights Articles

The Human Rights Council has reaffirmed in its resolution on the promotion, protection and enjoyment of human rights on the internet that human rights apply online as much as they do offline.[i] From the perspective of human rights, which can be impacted by the wrongful use of online and AI-enabled technologies, the rights to privacy, freedom of thought and expression, and non-discrimination are often identified as the most affected rights. Most human rights are not absolute and can be limited if certain requirements are met. In general, limitations to human rights provided in the International Covenant on Civil and Political Rights (ICCPR), including privacy, freedom of expression and non-discrimination, are allowed when such limitations are legally established in the law and are necessary and proportionate to achieve a legitimate aim.[ii]

Any organizational strategy to online monitoring should first and foremost be based on the core human rights principles orienting action to address hate speech, with the particular salience of Universal Declaration of Human Rights (UDHR) Article 19 and ICCPR Article 19:

Article 12, UDHR on the right to privacy, including non-interference in correspondence. Everyone has the right to the protection of the law against such interference or attacks.

Article 18, UDHR on freedom of thought, conscience and religion, including the right to manifest those beliefs through teaching, worship, and practice.

Article 19, UDHR on freedom of opinion and expression, including the right to communicate and share ideas through any media.

Article 16, UDHR on the right to education, including education for human rights and inter-cultural understanding.

Article 27, UDHR on the right to participate in the cultural life of the community, including the arts and science.

Article 17, ICCPR on the right to privacy and protection of that privacy by law "*no one shall be subjected to arbitrary or unlawful interference with his [or her] privacy, family, home or correspondence, nor to unlawful attacks on his honour and reputation*";"*everyone has the right to the protection of the law against such interference or attacks*." (Article 17(2)).

Article 9, ICCPR, on the right to freedom of opinion without interference, and freedom of expression, noting that:

> *this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice. The exercise of this right 'may be subject to certain restrictions, but these shall only be such as are provided by law and are necessary: (a) For respect of the rights or reputations of others; (b) For the protection of national security or of public order (ordre public), or of public health or morals'.*

Article 9, EU GDPR prohibits processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, subject to certain exceptions (see below). It also obliges all digital businesses to protect user privacy. There are exceptions to this provision, if:

- Article 9(a) – the data subject has given explicit consent;
- Article 9(c) – the data is necessary to protect the vital interests of the data subject;
- Article 9(e) – the data has already been made public;
- Article 9(g) – processing is necessary for reasons of substantial public interest.

Article 8, European Convention on Human Rights (ECHR) protects the right to respect for a person's private and family life, his home, and his correspondence. However, the right to respect for private life is a qualified right. Under European Court jurisprudence, an interference does not violate this right if it is in accordance with the law (i.e. based on a provision of domestic law that is accessible to the person concerned and foreseeable as to its effects); pursues a legitimate aim enumerated in Art. 8(2);[iii] and is necessary in a democratic society (i.e. the interference corresponds to a pressing social need and, in particular, is proportionate to the legitimate aim pursued).

Notes:

i Human Rights Council (29 June 2012). Resolution 20/...: The promotion, protection and enjoyment of human rights on the internet. A/ HRC/20/L.13. Available here.

ii See, for example, McKendrick, K. (August 2019). Artificial Intelligence Prediction and Counterterrorism, Chatham House. Available here;
Ganor, B. (2019). Artificial or Human: A New Era of Counterterrorism Intelligence. Studies in Conflict & Terrorism. UNICRI and UNCCT, New York.

iii American Association for the International Commission of Jurists (1984). Siracusa Principles on the Limitation and Derogation of Provisions in the International Covenant on Civil and Political Rights Annex. E/CN.4/1984/4 Available here.

ANNEX 2:
Example of a Project Risk Analysis

Used for the Digital Lighthouse Initiative, '*Applying Big Data and AI in the context of Hate Speech across Social Media*' (Regional Bureau for Arab States). *See case study for project description on page 35.*

CONSTRAINT (INTERNAL) / RISK (EXTERNAL)	DESCRIPTION	MITIGATION / STRATEGY
Risk (external)	**Reliability and integrity of big data sets** *Probability*: **Medium** *Impact*: **High**	1. Selection of known and recognized data provider. (Check with global market research companies.) 2. Clarify scope and limitations of datasets used. 3. Onboard data scientist to help shape the requirements for selection
Risk (external)	**Application of non-relevant conceptual framework and Big Queries** *Probability*: **High** *Impact*: **High**	1. Reach out to and partner with similar research projects by leading academic institutions for assistance. 2. Validate framework with national/local experts.
Risk (external)	**Varieties of Arabic language dialects** *Probability*: **Medium** *Impact*: **Medium**	1. Application of machine learning to refine the query results. 2. Creative and forward-thinking researchers from all disciplines to work together and identify nuances.
Risk (external)	**Reliability of AI and machine learning capability** *Probability*: **Medium** *Impact*: **High**	1. Selection of known and recognized AI experts (check within partner network on recommendation). 2. Clarify scope and limitations of technology used.
Risk (external)	**Reputational risk to engage on hate speech discussions** *Probability*: **Low** *Impact*: **High**	1. Involve legal experts to mitigate any grounds for misuse constitutional violation. 2. Chose countries.
Risk (external)	**Civil lawsuit filed by a hate group against UNDP citing 'free speech' based on constitutional grounds** *Probability*: **Low** *Impact*: **High**	1. Consultation of external experts to build capacity through experience internally.
Constraints (internal)	**Limited expertise on data / AI / hate speech** *Probability*: **High** *Impact*: **High**	1. Consultation of external experts to build capacity through experience internally.
Constraints (internal)	**Project is not covering all countries in the region** *Probability*: **High** *Impact*: **Low**	1. Identify additional countries to be launched after successful pilot 2. Communicate launch plan for additional countries in the region.
Constraints (internal)	**Long procurement process might challenge timeline** *Probability*: **High** *Impact*: **High**	1. Mitigate long procurement process through involvement of executive leadership.

Source: *UNDP Regional Bureau for Arab States. Digital Lighthouse Initiative. Applying Big Data and AI in the context of Hate Speech across Social Media. Available here.*

ANNEX 3:
Key Resources Highlighted in this Guidance Note

TOPIC	TITLE	DESCRIPTION	LINK
Violent extremist use of the internet	Deciphering the Siren Call of Militant Islamist Propaganda: Meaning, Credibility & Behavioural Change	This paper is particularly useful to gain an introduction to how online violent extremism (VE) messaging has the potential to leverage psychosocial forces and strategic factors to: (i) provide its supporters with a *system of meaning* that shapes how they perceive the world; (ii) demonstrate that it is a credible source of information and authority; and (iii) deploy pertinent behavioural levers designed to compel its audiences to legitimize and engage in violence.	icct.nl/publication/deciphering-the-siren-call-of-militant-islamist-propaganda-meaning-credibility-behavioural-change/
Risk management for preventing violent extremism (PVE)	Risk Management for Preventing Violent Extremism (PVE) Programmes - Guidance Note for Practitioners Supplementary Note to Support Revision of Risk Management Strategies In Context of Covid-19	This paper and supplementary note for COVID-19 risk management highlights the contextual, programmatic and institutional risks associated specifically with working on PVE programmes, drawing attention to the ways in which a context-specific, conflict sensitive, 'do no harm' and human rights-based approach can help to mitigate many of these risks and improve the effectiveness and efficiency of our programmes. From this solid basis, a risk assessment including more specialized potential risks related to using online data can be formed.	www.pvetoolkit.org/media/1178/josie-kaye-and-giordano-segneri-2019-risk-management-for-pve-programmes-guidance-for-practitioners.pdf www.pvetoolkit.org/media/1218/supplementary-note-for-covid-19-risk-management-for-pve-programmes-draft-270420-open.pdf
Guidance on implementing action on hate speech	United Nations Strategy and Plan of Action on Hate Speech Detailed Guidance on Implementation for United Nations Field Presences	This Guidance provides detailed information on how to implement the 13 commitments (related to both online and offline-based hate speech) set out in the United Nations Strategy and Plan of Action on Hate Speech and options for action that United Nations staff can take in field contexts, guided by the broad vision of prevention, and building on good practices from within the United Nations system as well as from Member States, civil society and other stakeholders.	www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20PoA%20on%20Hate%20Speech_Guidance%20on%20Addressing%20in%20field.pdf
Ethical guidelines for online research	Internet Research: Ethical Guidelines 1, 2 and 3	These documents offer a general structure for ethical online-based analysis, designed to help identify the ethically-relevant issues and questions, together with additional suggestions for how to begin to analyse and address these challenges in more detail. This general structure is provided as a guide for developing more extensive analyses of specific issues, both current and future.	aoir.org/ethics/
Privacy and the internet	The Right to Privacy In the Digital Age	The report provides guidance on how to address some of the pressing challenges that the right to privacy faces in the digital age. It provides a brief overview of the international legal framework and includes a discussion of the most significant current trends. It then turns to the obligations of States and the responsibility of business enterprises, including a discussion of adequate safeguards and oversight.	undocs.org/A/HRC/39/29

Data Principles	**Principles for Digital Development and Data Principles for UNDP**	The Principles for Digital Development is intended as a set of living guidance intended to help practitioners succeed in applying digital technologies to development programmes. The UNDP Data Principles demonstrate how this type of framework can be adapted to guide an organization in using data for development in line with its own mandate.	digitalprinciples.org data.undp.org/data-principles
Practical guidance on monitoring data for peacebuilding	**Social Media Analysis Toolkit**	This toolkit is a practical how-to guide for mediators and peacebuilders who want to conduct their own social media analysis, providing an overview of what is possible, a practical guide to a handful of technology tools, and suggestions on analysis methods.	rise.articulate.com/share/Sp41QVWIaGBvXvTKEEYY_Tb4FPGpBp92#/lessons/fytb6lqAC4f_OQs-3yAVjKPkd1D1EPI9
Artificial intelligence for P/CVE	**Countering Terrorism Online with Artificial Intelligence**	This report explores how AI can be used to combat the threat of terrorism online, and is relevant to many of the practical, ethical and rights-based issues of using AI tooling for the monitoring of VE trends and actors.	unicri.it/sites/default/files/2021-06/Countering%20Terrorism%20Online%20with%20AI%20-%20UNCCT-UNICRI%20Report.pdf
M&E for PVE	**Improving the Impact of the PVE Programming Toolkit**	This toolkit is designed for UNDP practitioners and partners who are working on programmes that are either specifically focused on PVE, or have PVE-relevant elements to them. It draws on best practice for design, monitoring and evaluation (M&E) in complex, conflict contexts adapting these for PVE programming. The toolkit includes modules, processes and approaches as well as an indicator bank that can be used within UNDP, with national and community-level partners, and as part of a capacity-building approach around monitoring. It can help set a framework for M&E for online-based projects that can then be adapted.	www.pvetoolkit.org/improving-the-impact
Online and offline PVE programming and policy	**PVE Knowledge and Research Portal**	UNDP's online research portal for publications on PVE can help inspire ideas on online- and offline-based projects towards PVE, such as which online data collection could be used to inform.	pveportal.org


UN
DP