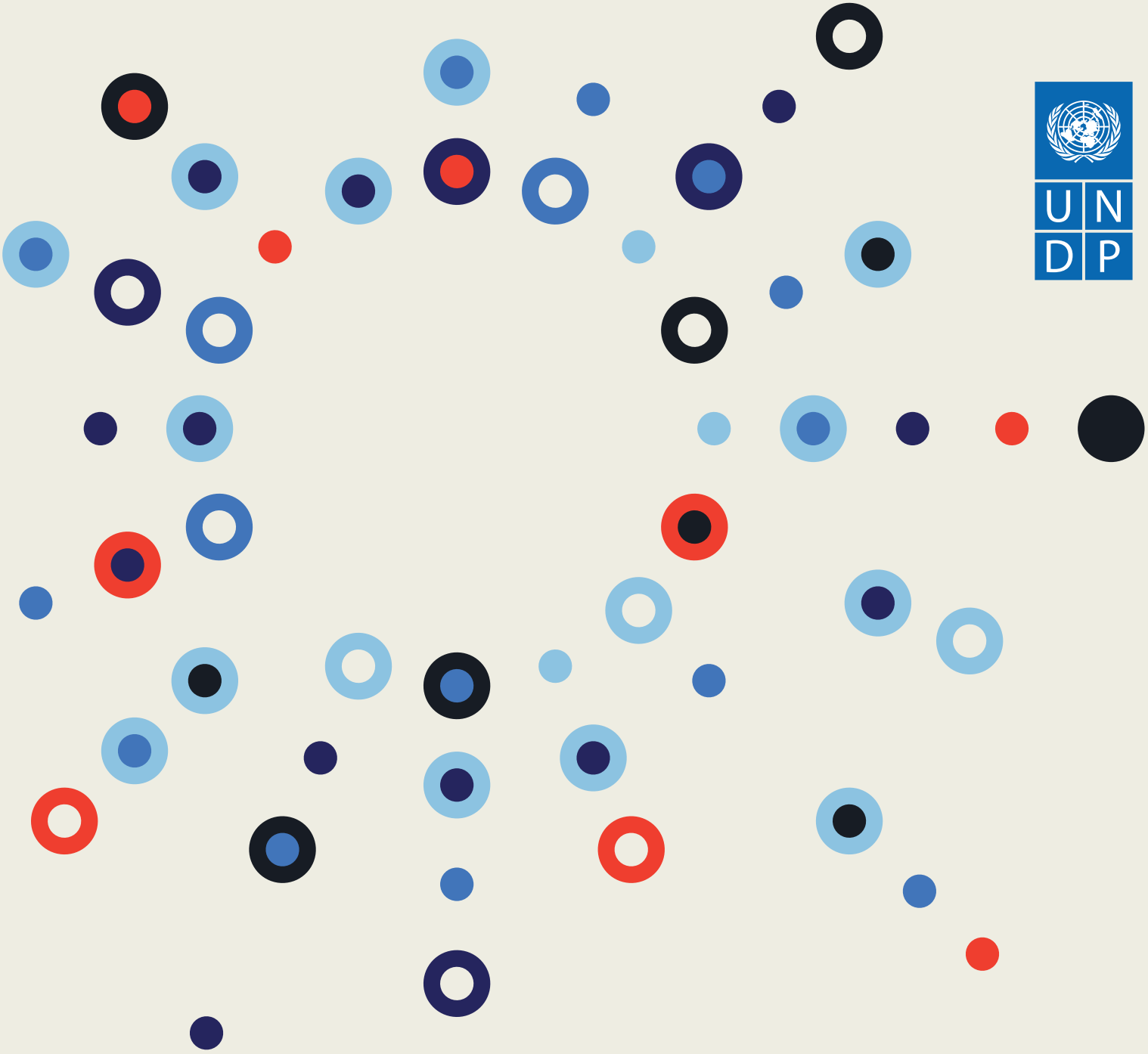


POLICY BRIEF

From Pilots towards Policies:

Utilizing online data for preventing violent extremism and addressing hate speech

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

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 policy brief 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 guidance of Samuel Rizk (PhD), the development of the policy brief was supported by the lead author Angharad Devereux and contributing author Gitte Nordentoft. The Team is grateful to the UNDP Global Policy Network (GPN) of PVE practitioners, Regional Hubs and Country Offices, UNDP Chief Digital Office, UNDP Crisis Bureau CPPRI/Conflict Prevention and Peacebuilding Team and Oslo Governance Centre for their valuable inputs to consultation process, including to the concept note, initial draft and contributions within a feedback event. The Team would like to thank the Cyber Threats Research Centre of Swansea University, particularly Joe Whittaker and Professor Stuart Macdonald, for their peer-review of the research process and document from conception to completion. The authors would also like to thank GIFCT, Tech Against Terrorism, Moonshot, Codetekt, Koe Koe Tech, RUPANTAR and RIWI, for their contributions of expertise, including those representatives who contributed to the event held by UNDP and the EU Commission to gather information within UNOCT Counter Terrorism Week, titled 'The Opportunities and Challenges Presented by Online and AI Tools for the Prevention of Violent Extremism.'

Design and layout: Benussi & the Fish

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

Introduction

As the Internet has expanded in reach to more individuals in increasingly mobile ways since the advent of the web 2.0, so too have opportunities for those wishing to use this globalised networking architecture for harm, including violent extremism (VE).[1] With an increase in online information and data related to violent extremist use of the internet does, however, come the opportunity to gain insight into Violent Extremist Groups' (VEGs) narratives and receptive audience as well as potential driving factors (e.g. grievances and current affairs) that may be utilized by VEGs to sow distrust in societies and institutions.

The drivers of violent extremism are contextual and the factors that have tipped dissatisfaction and radicalisation into violent extremist behaviour vary on a case-by-case basis.[2] The impact of the COVID-19 pandemic and corresponding responses have potentially worsened many of the drivers of violent extremism,[3] therefore heightening the likelihood of susceptibility to violent extremist content. This has been compounded by the widespread, yet uneven,[4] digital surge that has stemmed from the pandemic,[5] allowing for increased exposure to violent extremist material and hate speech online and providing the foundations for a breeding ground of conspiracy theories, disinformation and extremism.[6]

Online data and emerging technology, including artificial intelligence (AI) tooling, offer the opportunity to heighten understanding of the impacts of events or issues, such as COVID-19, on social sentiment and help pre-empt tactics VE groups may use to manipulate these. However, these pioneering initiatives are linked with major structural, methodological, technical, practical, ethical and human rights-based challenges, from the resources to the responsible partnerships required, which are heightened by limited subject-specific guidance, including on risk management and M&E.

In this context, this Policy Brief explores the potential of utilizing resourceful, efficient, ethical and rights-based data-driven methods to inform Prevention of Violent Extremism (PVE) programming, aligning with the 2021 General Assembly Resolution 75/291 calling for "developing an accurate understanding of how terrorists motivate others to terrorist acts or recruit them, and develop the most effective means to counter terrorist propaganda, incitement and recruitment, including through the Internet and other information and communications technologies".[7]

UNDP's approach to Prevention of Violent Extremism (PVE)

In alignment with the Secretary General's PVE Plan of Action and the UN Global Counter-Terrorism Strategy, UNDP supports more than 40 countries to prevent violent extremism. UNDP's work on PVE address two interlinked challenges: (1) the phenomenon of violent extremism, using a development and peacebuilding approach firmly grounded within human rights principles, and (2) the need to govern increasingly diverse and multicultural societies, which requires attention to institutions, people's identities, including political and religious ideologies, and the promotion of human rights-based approaches. The focus of UNDP's preventive approach is to look at the relationship between peaceful societies and inclusive development; rule of law and human rights; anti-corruption, good governance, civic engagement and political participation; and to address the inequalities that fuel radicalisation that can lead to violent extremism. Projects that utilize online data for PVE may therefore concentrate on collecting and analysing data that pertain to any or all of these aspects with a focus on how they relate to violent extremism in a given context.

Processes for utilizing online data and AI application

While there is still more to be understood, there is general agreement among researchers that the internet provides opportunities for radicalisation.[8] VE groups are found to utilize online spaces for recruitment, propagandising, development of community, psychological warfare, planning, information sharing, networking and financing in different ways depending on the group, ideology and individual factors.[9]

Building VE identity and legitimacy

VE group identity is created through an escalating process of linking concepts of crisis or threat to blame and marginalise the out-group. The out-group is usually deemed as those outside of the VE group (the in-group) and is often centred around one particularly condemned group and can extend to the active or tacit supporters of that group.[10] Legitimacy and shared identity is built through offering the solution to this perceived crisis within the in-group.[11] This crisis perception encourages a perceived vulnerability of the in-group and automatic, non-deliberative thinking, which in turn lends itself to encouraging individuals to believe and possibly act on views that sway far from normality.[12] This is strengthened by the system of meaning (new lenses by which to interpret the world in line with group norms) that extremists create for members through use of propaganda.[13]

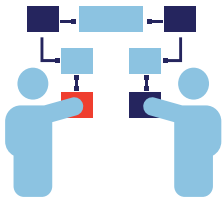
VE groups will directly or indirectly adopt an 'us versus them' narrative, which lends individuals to being more willing to fight or die for their group. Such propaganda can take the form of formal or informal text, image, or video and will often use and manipulate historical or current events to legitimise the argument.[14] In this way, the internet offers a new outlet and communal stage for violence towards the out-group. VE groups utilize the internet to flood individuals with this material,[15] whilst also using techniques to make internet users feel that they belong to an exclusive group, for example by using invite only chat rooms[16] and open links to access hidden materials online,[17] which may foster an in-group mentality.

Online manifestations of violence and hate against women

Research from **UNDP Pakistan** focused on attitudes towards women in online spaces, analysing social media posts to understand the gendered online expression of current affairs. The research shows that as violence against women persists, it also mutates into new forms online and revitalises the acceptability for gender tropes, policing transgressions and subverting the gains that access to democratized media platforms potentially promised women.[18] The research points out the particularly harmful effect this has on normalising misogyny in young internet users, which is a known contributing factor to increased levels of violence in society, including violent extremism.[19]

Preventative efforts aim to understand the driving factors of violent extremism in a given context and contribute to informed policy and programming to prevent the appeal of VE spaces online in the first instance, whilst in turn supporting institutional structures. Informed by UNDP pilot programming, the data cycle outlined below identifies the main elements of PVE programming that utilizes online data.

Data Cycle

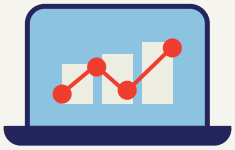


a) Objective, methodology design and risk assessment

To utilize 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.

.

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.

d) Application

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

Illustrative example of the data cycle

UNDP Bangladesh has devised a project of which the objectives are to understand audiences of Bangladeshis attracted to extremist narratives and to form a better understanding of 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 developed alongside the partner, SecDev, based on a public health and expert-driven social science approach. Collection is undertaken through use of manual and automated scrapers. Analysis is organised and communicated into regular reports. A dashboard is being developed in order to organise this in line with project aims. The live insights gained provide ongoing programming support for both online and offline PVE initiatives by applying understanding gained of the range of violent extremist and exclusionary rhetoric in Bangladeshi cyberspace. The project's methodology is assessed and tested on target audiences in order to inform the next round of reporting and to ensure objectives are being met. Data gained through this project has increased the reach and impact of pro-peace and tolerance narratives in target audiences, reduced engagement with negative VE content, and has increased multi-stakeholder national understanding of VE trends.

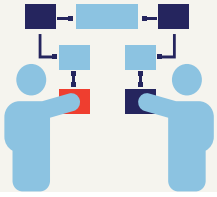
Throughout the cycle it is important to resist a reliance on online-based methods alone. Generally, there is a need for human-centred creativity and traditional forms of data collection to be integrated with data-driven approaches to PVE programming. This is particularly pressing as the nature of online data made available from social media platforms is that information individuals choose to share. Material intentionally shared for public display is not necessarily an accurate representation of one's true self but rather a purposefully constructed, digitally mediated identity, or digital avatar.[20] Data collection and analysis can provide insight on the gender, demographic profile, location and other measurable characteristics of followers when these users consent to sharing this data on public platforms and utilizing this data is authorised by the Terms of Service of the platform(s) in question.

Learning through adaptation

The **UNDP Digital Lighthouse** Initiative invests in research initiatives that act as a 'lighthouse' for similar programmes across the organisation. One such 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. The project has faced major challenges involving both the resources necessary to carry out a project involving big data, and AI, as well as the fact that the data made available by Twitter made it difficult to reach the stated objectives of the project. The project research team therefore reacted to this challenge by adapting the codebook used to label data from monitoring hate speech to 'negative speech' and much useful data was found in this 'grey' 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 utilized by this project.

Policy considerations throughout the data cycle

The data cycle provides an overview of the process of utilizing data and helps dissect existing challenges to each phase which require further policy consideration and systematisation.



a) Objective, methodology design and risk assessment of online data

Each stage of the data cycle should be analysed and adequate risk assurance and mitigation measures identified. One major risk of utilizing online data for PVE is the misuse of data by government actors. Human rights record of government recipients of data and the potential of misuse of data collected should be incorporated into risk assessment. The risk of misuse of data to support political agendas is compounded by the lack of an internationally accepted definition of terrorism.[21] The UN Security Resolution 1566[22] and the definition of terrorism crafted by the Special Rapporteur's office on Terrorism and Human Rights[23] do attempt to set boundaries but due to a lack of a multilateral treaty, many national definitions of terrorism are wide and vague enough that much of what is targeted online and offline is legitimately protected by international law. Tech companies have tried to set definitions but these are often self-created in processes outside of the multilateral context, where external actors, such as subject-specific experts and civil society, have little consistent access. Indeed, even when definitions are clear, these can be manipulated to designate actors or material as terrorist or not. The lack of definitional frameworks in this area also impacts the collection stage of the data cycle as an international definitional foundation to direct the collection of data related to terrorism does not exist, therefore necessitating subjectivity at this stage. Terrorist designation lists do not solve this problem due to challenges such as the processes by which the lists are constructed and the due process rights of members of these lists.

Online data sources can range from blogs to online news outlets, though social media platforms are the most popular source of information due to unparalleled global engagement. However, large social media companies operate on for-profit business models which encourage maximum engagement. This could be in conflict with the ultimate objective of PVE, which includes preventing patterns of hate speech and dehumanisation and building cohesive communities. Polarisation of society can be encouraged by recommendation algorithms used by social media platforms that encourage agreeable information due to the increased likelihood of clicking on such information.[24 25] In this way, these algorithms can also encourage increasingly extreme information, creating incentives to shock and intrigue for clickbait. On the most harmful end of the scale is the fact that recommendation systems have been shown to have the potential to increasingly feed an initial interest in extreme material[26] and aid the creation of alternative news networks.[27] Algorithms have become more apt at finding 'rabbit holes'[28] or 'filter bubbles' which encourage the bypassing of thoughtful consideration by dramatically amplifying confirmation bias.[29] This is strengthened 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. The constant stream of information also encourages 'system 1 thinking' ('thinking fast'), conducive to violent extremist groups' aims to create communities based on hatred of the out-group,

'operating automatically and quickly, with little or no effort and no sense of voluntary control'.[30] This is in comparison to 'system 2 thinking', or 'thinking slow', which 'allocates attention to the effortful mental activities that demand it, [and] is associated with the subjective experience of agency, choice, and concentration.'[31]

In the multi-layered policy landscape that has evolved from increased online presence of VE groups, responsibility is often passed between private, regional, and national governmental actors, resulting in a lack of clear lines of responsibility and transparency of guiding principles, rules, and regulations. Unlike national laws that are limited by geographic borders, Terms of Service agreements apply to platforms' services on a global scale. Legal attempts to regulate tech companies can be reactionary and, depending on public pressure, can create demands for security actors to access data, to enhance speedy content removal, to protect freedom of expression online and/or increase privacy standards of citizens, demands which can often be competing and create difficulty for tech companies to meet without imposing on freedom of information and/or privacy.[32] The Global Internet Forum to Counter Terrorism (GIFCT) has been developed with the purpose to engage in cross-industry technical collaboration, standard setting and guidance in this area, governed by four founding tech companies and an independent advisory board.[33] However, organisations like this alone cannot be depended on as a silver bullet to the issue of fragmented policy and practice due to the fact that they lack the transparency and legality of multilateral spaces, i.e. an intergovernmental entity bound by formal international law.[34] Clarity, understanding and transparency is 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 of one another.[35]

First and foremost, to policy or programming-based efforts to preventing and countering VE narratives and addressing hate speech, the UN and Member States must follow the overarching principles of international human rights law as enshrined in international instruments, such as the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights, as well as the UN Guiding Principles on Business and Human Rights and the Rabat Plan of Action. These principles should be considered by all stakeholders and translated by organisations into actionable frameworks for utilizing data in a human rights-compliant manner. As an example of operationalising broader human rights law, The UN has developed key frameworks to help international policymakers and practitioners navigate the application of human rights in this area. For example, The UN Strategy and Plan of Action on Hate Speech[36] sets out the objectives of enhancing UN efforts to address root causes and drivers of hate speech and enabling effective responses to the impact of hate speech on societies. This Plan created a clear strategy and platform to meet these objectives, in turn helping create clarity on those most relevant stakeholders to convene in order to reach these. The Secretary General's UN Data Strategy[37] 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 formed a platform to collaboratively design and implement data policies that advance the responsible human-rights-based use of data.

b) Collection of online data

Governmental efforts (national and regional) to counter VE use of the internet have largely rested on passing responsibility to social media companies to remove VE content in a timely manner, with heavy fines in cases of non-compliance. This can encourage overzealous removal to meet goals, place unrealistic pressures on small platforms, and ignore more systematic issues.[38] These approaches to counter violent extremist utilization of the internet have rested on the intent of removing the ability of VE groups to spread propaganda, recruit and coordinate online. These approaches are therefore largely defensive, reacting to VE use of the internet, rather than proactively preventing this.[39] The positive result of content removal as a whole is that major platforms become more difficult to utilize by VE groups, meaning audiences for propaganda are lessened, however VE groups/individuals can become more difficult to pre-empt by law enforcement as the dark web/harder to reach platforms are increasingly utilized.[40] Content removal practices also create displacement effects that change the nature of VE activity online and data available for monitoring as VE groups and individuals attempt to manoeuvre detection by creating content that falls outside of platform Terms of Service either by nature of its content or by technologically evading content detection tooling.[41] As VE content becomes more adaptable to these practices and less clearly identifiable, tackling and/or utilizing 'harmful yet legal' content online necessarily relies on a subjective interpretation of harm which is particularly difficult to implement at scale without negatively impacting freedom of expression.

AI tools can be used to enhance the volume of data collected, organised 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. To be classed as 'fair', algorithmic decisions must not create a discriminatory or unjust impact on the end-users. Whilst the technology behind AI tooling will be a neutral statistical and mathematical process, AI can amplify existing biases of societies, or within the teams developing the tool, when trained with biased datasets, resulting in incidents of automated solutions discriminating against individuals, groups, and/or communities on prohibited grounds.[42] Linked to this is the need for transparency and ability to explain algorithmic decisions and actions taken on the basis of such results.[43] Limitations of accurate and ethical data collection and amplification can also stem from the extent and nature of available data, including potential gender, ethnicity, and/or class-based misrepresentation. A lack of available data on particular VE or 'at-risk' groups, or inability of tooling to pick up the language of such groups, may also create disproportionate focus and misleading results in a given context. To overcome this, regulation and monitoring of access and use of AI tooling is needed. For specific interventions, diverse and multi-sectoral teams with contextual knowledge and technical expertise to design and consult with can help enhance the accuracy of this process as well.

Good practice of tool development processes

The **Terrorist Content Analytics Platform (TCAP)** is a secure online tool that automates the detection and analysis of verified terrorist content on the internet. To ensure accuracy and accountability, the TCAP development began with a multilateral consultation process on the subject of ensuring transparency by design, seeking insights from civil society groups, tech companies, and academics.[44] This consultation is also built into the design of the tool in order to ensure appropriate mechanisms to uphold human rights, freedom of expression, transparency, and accountability. From this, accessible monthly virtual 'office hours' are held, an academic advisory board is being formed, and a reporting and appeals mechanism for sharing of views on the classification of content has been offered. As Tech Against Terrorism has stated, 'such systems at the structural and systematic level need to support these vital checks and balances – these mechanisms cannot be an afterthought'.[45]

c) Analysis of data

It follows that countries with lesser internet penetration rates and access to platforms where open and free expression is permitted will be able to offer a smaller volume of data to be utilized for the purpose of PVE. The possibility of connection 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. Certain languages and dialects are also underrepresented in monitoring tooling available in these contexts due to potential lessened demand, expertise and funding available. Local capacity and digital literacy of partners and government actors can be further limiting consequences of lack of online access, particularly in the least developed and conflict-affected areas, which can create a need for customised, and therefore more costly and complex solutions.[46] Therefore, an overreliance on online data would naturally act as a detriment to countries challenged by the digital divide. Furthermore, social media analysis may also overrepresent the views of particular demographic groups (particularly related to class and gender) within society, which may use certain platforms more than others and have greater access to devices.[47]

A focus on VE utilization of the internet, as well as online policy and practice-based responses, should not misplace the issue of 'radicalisation' to the online sphere alone and perpetuate a false dichotomy between the online and offline, as in daily life across the globe now, the online and offline spheres are deeply interlinked.[48] Research suggests that offline interactions are key to radicalisation, which may act as a counterweight to growing online activity.[49] Therefore, efforts to understand and react to what happens online should be considered as only one part of the larger analysis. This highlights the need for integrated government approaches and frameworks to PVE (such as PVE National Action Plans) that address both offline and online spaces.

d) Application of collected data

A development approach to PVE rests on tackling root causes such as governance grievances (e.g. corruption and inadequate service delivery), lack of (e.g. economic and educational) opportunities, misogyny and domestic violence, and poor mental health, in order to develop long-term, holistic approaches to PVE. It aims to lessen the appeal of VE rhetoric and communities offline and online by heightening individual and societal opportunity and sense of meaning.[50] To adequately address the drivers of violent extremism, online-based PVE programming must be enhanced holistically, including projects that tackle the appeal of VE content online, promoting critical analysis and alternative, positive messaging. Thoughtfully designed digital citizenship programmes can inspire critical and technical ability to confidently navigate the online sphere and recognise how VE narratives may present themselves and aim to manipulate users' attention.[51]However, access to systematic digital-based capacity building is uneven and curriculums often do not cover topics relevant to the modern threat landscape such as the purpose and potential effects of disinformation, recommendation algorithms, addiction technology and exclusionary, unrealistic norms perpetuated by social media. Additionally, positive/alternative messaging can lack long term-planning, links to offline initiatives, credibility, targeting and meaningful monitoring and evaluation beyond statistics of reach.[52] These are all factors which government authorities at different levels must consider in education planning, civic education and awareness raising.

e) Monitoring and Evaluation

Strong systems for monitoring and evaluating efforts for utilization of online data is essential due to the relatively novel methods and technologies applied. There is a need for existing M&E approaches to be tailored and updated, utilizing innovative methods such as behavioural insights, integrated data assessments and A-B testing to ascertain what works well to achieve stated objectives.[53] When enacted alongside traditional data collection methods, such as surveys and literature reviews, a better understanding of factors such as target audience, messaging, medium and platforms for future projects can be obtained. This is challenging due to the fact that PVE efforts navigate in landscapes which are not fully understood, and which are continuously evolving. A growing awareness of potential risks through piloting programming allows us to begin to delineate these into more systematic and tailored M&E frameworks, in order to ultimately offer greater guidance to those undertaking these processes on the ground which will further create an evidence base much needed for policy makers. These learnings can also be used to pinpoint gaps in risk assessment procedures and modify these accordingly.

Encouraging multistakeholder engagement to protect and advance

Solutions to the identified policy and programmatic challenges should be found in multistakeholder collaborations which include civil society and tech platforms, governments, and academia to leverage new expertise, technology and capacity and hence enhance quality, efficiency, legitimacy, and relevance of interventions. However, engaging in partnerships should never be seen as a passing of responsibility and rigorous assurance of stakeholder ethical and human rights standards must come first and foremost. Part of this is ensuring that stated methodologies are transparent, as are employment standards of those working for the organisation in question, and data collection, storage and preservation practices are systematically risk assessed and human rights compliant.

Civil Society Organisations (CSOs) offer contextual expertise which is important when analysing online discussions necessitating local knowledge, and fluency in local languages and dialects. Granular local knowledge can help avoid some of the biases amplified by machine-learning algorithms. Localised actors can play a vital role in highlighting the lived experience within the communities they serve of the dangers of private companies implementing new algorithms, as well as advocating this to the audience of policymakers responsible for governing this space through regulation. An inclusive and diverse range of actors involved in policy and programme development will always ultimately strengthen outcomes and create opportunities for local capacity building and engagement. This therefore aids the creation of thriving and resilient communities to VE groups which aim to make individuals feel marginalised to the fringes. It is vital to empower the capacity of CSO actors in ensuring fairness, accountability, and transparency of automated systems in the future, particularly as algorithmic decision processes are increasingly impacting everyday life.[54] CSO's rights and wellbeing must be ensured through appropriate policies and practices at all levels.

Relying on the data developed by social media platforms presents an enduring problem as private companies do not have the same obligations as states or individuals when it comes to adhering to international human rights mandates.

Due to a lack of formal governmental regulatory oversight of tech companies, platforms that are regulating the access to use of their services are by and large the standard setter, the enforcer and the arbiter of policy and practice. In light of the lack of 'hard law' mechanisms, oversight is often self-created by tech companies,[55] a fact that is made more relevant when one considers their role in access over and to information, freedom of opinion and expression, freedom of assembly, and public interest discourse, including in the context of health or election space.

Transparency on human rights compliance of these platforms' Terms of Service, including user consent for data usage, and redress mechanisms, is needed and meaningful external, including democratic oversight, is a crucial consideration when utilizing the data of such platforms.

Partnership with development actors can act as a positive guise of human rights compliance for tech companies and consideration of whether the business model is in fact at odds with efforts to prevent divisive, harmful, and violent content should be incorporated into risk assessments as well as in policy development.

Partnerships can be used however to foster and advocate for greater human rights compliance of organisations, using the tools and expertise of respective stakeholders to strengthen the approach of all towards preventing violent extremism.

Recommendations towards stronger policies

To summarise, this Policy Brief aims to gather lessons learnt from UNDP piloting in the utilization of online data for PVE to advocate for the development of policy that encourages consistently human rights-compliant, data driven PVE programming to complement traditional data collection and application methods. The Policy Brief encourages stakeholders to execute this in an informed, transparent and responsible manner that builds the capacity of the international community of development actors to prevent violent extremism.

To enact this, this Policy Brief recommends the following policy and programming-based actions for policy/decision makers:

1. **Addressing root causes of violent extremism.**
   Any development of PVE policy and programming should start from an understanding of *why* individuals do or do not turn to online VE material in the first instance. Therefore, policy and programming must address root causes by taking both online and offline approaches of data collection and application to building meaningful existences.

2. **Strengthening and ensuring transparency of definitions.**
   Application of definitions related to 'hate speech' and 'terrorism/terrorist group' must comply with Human Rights Frameworks in order to lessen subjectivity and mitigate the risk of misuse of findings. Subjective interpretation of harm should be discouraged wherever possible, as these are almost impossible to operationalise at scale without negatively impacting freedom of expression. Entities should clearly state the definition of derivation, including of the substance of what it means to 'encourage' violent extremism.[56] To aid all stakeholders, Member States should work towards devising their own contextualised definitions of hate speech in order to reduce arbitrary restrictions online, strengthening these practices by ensuring that governments fulfil the three-fold test of legality, necessity, and proportionality in those instances when 'hate speech' is punishable by law.

3. **Assessing and guiding tech partnerships.**
   Engagement with the big tech companies should be recognised as having unique implications for risk assessment processes, which must include consideration of Terms of Service, content moderation practices, redress mechanisms, algorithmic transparency, and business models. Therefore, clear guidance around operations and processes when procuring these companies should be provided for practitioners. States should uphold their responsibility to ensure that businesses that are operating in their jurisdiction respect human rights.

4. **Developing positive online ecosystems and information skills development.**
   Information literacy and critical thinking curriculums and programmes should be enacted, based on the current opportunities and threats posed by the internet, incorporating both technical competency and critical ability, to increase resilience and lessen the strategic interest of the internet for VE groups in the first instance. A range of online-based programmes created to incubate positive narratives online can complement these efforts to build resilience online.

5. **Building local capacities.**
   Local actors, particularly CSOs, should be empowered to be partners of choice in order to ensure methods are built on sound contextual understanding, reducing potential bias in methodologies. However, the protection of partners' mental and physical wellbeing should always be prioritised from the design stage.

6. Strengthening risk assessment and M&E.

Organisations should develop risk assessment methods that clearly state the ethical and human-rights considerations of projects that utilize online data and AI for PVE by applying and expanding on existing risk mitigation measures and tools such as the UNDP PVE Risk Guidance [57]. Investing in gaining an in-depth understanding of the potential types of risks posed by the utilization of online data and AI for PVE is necessary, gathered through literature, policy, and practice review. These should be strengthened by proper investment into meaningful M&E frameworks which encourage consistent monitoring of what could cause harm and evaluation of how to mitigate these factors in order to more effectively utilize online data. Whilst difficult to form a 'one-size-fits-all' static solution in this area due to the variation of risk depending on context and rapidly evolving threat, technological and policy landscape, the aim should ultimately be to reduce the burden on practitioners undertaking these projects in forming risk assessment processes themselves.

7. Prioritising digital transformation.

Organisations should recognise and react to the growing prevalence of data in any organisation that strategically and systematically works in the sphere of PVE by building staff capacity to utilize online data and AI for PVE in an effective and informed manner, understanding both the opportunity and the limitations involved. What digital transformation means to a particular organisation must be first identified and then translated to staff in a clearly relevant and applicable manner. To manoeuvre the complex policy landscape, people's rights should always be based as the ultimate starting point for strategic frameworks.[58] To facilitate this, organisations must create a strategy that is guided by principles that encompass human rights and ethical consideration.

Endnotes

1 Von Behr, I. Radicalisation in the Digital Era, RAND Corporation, 2013.

2 UNDP, 'Preventing Violent Extremism Through Promoting Inclusive Development, Tolerance and Respect for Diversity', 2016.

3 Avis, W. 'The COVID-19 Pandemic And Response On Violent Extremist Recruitment And Radicalisation', K4D, 2020. Specific factors which may have been worsened include reductions in civic space (CIVICUS, 'Freedom of Expression and the COVID-19 Pandemic: A Snapshot of Restrictions and Attacks', 2021) ; misuse of terrorism laws (OHCHR, 'UN Expert Says Misuse of Terrorism Laws During Conflict Risks Worsening Situation', 2020); poorer mental health among young people in particular (UNICEF, 'The Impact Of COVID-19 On The Mental Health of Adolescents and Youth', 2020). The pandemic has also increased poverty rates, especially in regions affected by violent extremism (IMF, 'World Economic Outlook, April 2020: The Great Lockdown', 2021).

4 Due to the global digital divide

5 De, R. 'Impact of Digital Surge During Covid-19 Pandemic: A Viewpoint on Research and Practice', International Journal of Information Management, 2020.

6 UNDP, 'Responding To COVID-19 Information Pollution', 2020; Cox, K., 'COVID-19, Disinformation and Hateful Extremism', RAND Europe for the Commission for Countering Extremism, 2021; Commission for Countering Extremism, 'How Hateful Extremists Are Exploiting the Pandemic', 2020.

7 United Nations, Resolution Adopted by The General Assembly On 30 June 2021.

8 Aly, A. Introduction, Violent Extremism Online: New Perspectives on Terrorism and the Internet, 1st edn, Routledge 2018, p. 4.; European Commission, 'Prevention Of Radicalisation - Migration And Home Affairs, 2021; Borum, R. Psychology Of Terrorism, University of South Florida 2004.

9 Gill, P. 'What Are the Roles of the Internet in Terrorism?', VoxPol, 2015; Conway, M. 'Reality Bytes, PhD: 'Cyberterrorism and Terrorist Use of the Internet', Trinity College, Dublin, 2006.

10 Berger, J. M. Extremism, MIT Press, 2018.

11 Ingram, H. 'A Linkage-Based Approach to Combating Militant Islamist Propaganda: A Two-Tiered Framework for Practitioners', Terrorism and Counter-Terrorism Studies, 2016; Tajfel, H. 'Individuals and Groups in Social Psychology', British Journal of Social and Clinical Psychology, 1979; Berger J.M., 'Extremist Construction of Identity: How Escalating Demands for Legitimacy Shape and Define In-Group and Out-Group Dynamics', 2017.

12 Ingram H., 'Deciphering the Siren Call of Militant Islamist Propaganda: Meaning, Credibility & Behavioural Change' Terrorism and Counter-Terrorism Studies, 2016.

13 Ibid.

14 Ellemers, N. 'The Group Self', Science, 2012; Swann, W. 'Identity Fusion: The Interplay of Personal and Social Identities in Extreme Group Behavior', Journal of Personality and Social, 2009 Psychology; Waytz A., 'Motive Attribution Asymmetry for Love Vs. Hate Drives Intractable Conflict', Proceedings of the National Academy of Sciences, 2014; Stephen Reicher, S. Alexander Haslam and Rakshi Rath, 'Making A Virtue of Evil: A Five-Step Social Identity Model of The Development Of Collective Hate', Social and Personality Psychology Compass, 2008; Cikara, M., 'Their Pain Gives Us Pleasure: How Intergroup Dynamics Shape Empathic Failures and Counter-Empathic Responses', Journal of Experimental Social Psychology, 2014

15 Ali Fisher, 'Swarmcast: How Jihadist Networks Maintain a Persistent Online Presence' Perspectives on Terrorism, 2015.

16 Bloom, M, 'Navigating ISIS'S Preferred Platform: Telegram 1', Terrorism and Political Violence, 2019.

17 Weimann, G. 'Terrorist Migration to The Dark Web', Perspectives on Terrorism, 2016.

18 UNDP Pakistan, 'Masculinity and Online Abuse: Digital Discourses Targeting Women in Pakistan', 2021.

19 UNDP Oslo Governance Centre, 'Misogyny: The Extremist Gateway?', 2021; Johnston, M. 'Misogyny & Violent Extremism: Implications for Preventing Violent Extremism' UN Women Asia and Pacific, 2019.

20 Gündüz, U. 'The Effect of Social Media on Identity Construction', Mediterranean Journal of Social Sciences, 2017; Burkell, J. 'Facebook: Public Space, Or Private Space?' Information, Communication & Society, 2014; Uimonen, P. 'Visual Identity in Facebook', Visual Studies, 2013.

21 For a discussion of the challenges related to defining terrorism, see Stella Margariti, 'Defining International Terrorism to Protect Human Rights in The Context of Counter-Terrorism', Security and Human Rights, 2018.

22 UN Security Council Resolution 1566 (2004) on threats to international peace and security caused by terrorist acts.

23 UN General Assembly, 'Report Of The Special Rapporteur On The Promotion And Protection Of Human Rights And Fundamental Freedoms While Countering Terrorism, Martin Scheinin Ten Areas Of Best Practices In Countering Terrorism', 2010.

24 For a breakdown of the different types of algorithms used by social media companies see: GIFCT, 'Content-Sharing Algorithms, Processes, And Positive Interventions Working Group', 2021.

25 UN General Assembly, 'Report Of The Special Rapporteur On The Promotion And Protection Of Human Rights And Fundamental Freedoms While Countering Terrorism, Martin Scheinin Ten Areas Of Best Practices In Countering Terrorism', 2010.

26 Reed, A. 'Radical Filter Bubbles: Social Media Personalisation Algorithms and Extremist Content', Global Research Network on Terrorism and Technology, Paper No 8, 2019.

27 O'Callaghan, D. 'Down The (White) Rabbit Hole: The Extreme Right and Online Recommender Systems' Social Science Computer Review, 2014.

28 Pariser, E. The Filter Bubble: What the Internet is Hiding from You, Penguin, 2013.
29 Lindström, B. 'A Computational Reward Learning Account of Social Media Engagement' Nature Communications, 2021.
30 Ingram, H. 'Deciphering the Siren Call of Militant Islamist Propaganda: Meaning, Credibility & Behavioural Change' Terrorism and Counter-Terrorism Studies, 2016.
31 Kahneman, D. Thinking, Fast and Slow, Penguin, 2012.
32 Keats, D. 'Extremist Speech, Compelled Conformity, And Censorship Creep', SSRN Electronic Journal, 2017.
33 GIFCT, 'Resource Guide', 2021.
34 Douek, E. 'The Rise of Content Cartels', SSRN Electronic Journal, 2020.
35 Ní Aoláin, F. 'The Impact Of "Soft Law" And Informal Standard-Setting in the Area of Counter-Terrorism on Civil Society and Civic Space', The University of Minnesota, 2020.
36 United Nations, 'UN Strategy and Plan of Action on Hate Speech', 2019.
37 UNSG, 'UN Secretary-General's Data Strategy'.
38 Tech Against Terrorism, 'The Online Regulation Series: The Handbook', 2021; OHCHR, 'Facebook Sharing Buttontwitter Sharing Buttonlinkedin Sharing Button Moderating Online Content: Fighting Harm Or Silencing Dissent?'
39 For a discussion of a reliance on defensive tactics in the online CVE space, see Alastair Reed and Haroro Ingram, 'A Practical Guide to the First Rule of CTCVE Messaging', Europol, 2019.
40 Whittaker, J. 'The Online Behaviors of Islamic State Terrorists in The United States', Criminology & Public Policy, 2021. See pages 181/182 for a summary of research relating to this hypothesis.
41 Macdonald, S. 'Regulating Terrorist Content on Social Media: Automation and the Rule of Law', International Journal of Law in Context, 2019; OECD, 'Current Approaches to Terrorist and Violent Extremist Content Among the Global Top 50 Online Content-Sharing Services', 2020.
42 United Nations Office of Counter-Terrorism (UNOCT), 'Countering Terrorism Online with Artificial Intelligence', 2021.
43 Ibid.
44 Tech Against Terrorism, 'Conclusions from the Online Consultation Process for The Terrorist Content Analytics Platform (TCAP)', August 2020.
45 VoxPol, 'The Terrorist Content Analytics Platform and Transparency by Design', 2020.
46 United Nations Office of Counter-Terrorism (UNOCT), 'Countering Terrorism Online with Artificial Intelligence', 2021.
47 For example, men remain 21% more likely to be online than women, rising to 52% in the world's least developed countries. Blog by Iglesias, C. 'The Gender Gap in Internet Access: Using A Women-Centred Method', 2020.
48 Whittaker, J. 'Understanding the Online and Offline Dynamics of Terrorist Pathways'; Bishop, P. 'Response to The Online Harms White Paper', CYTREC/Swansea University, 2019, p. 2.
49 Whittaker, J. 'The Online Behaviors of Islamic State Terrorists in The United States', Criminology & Public Policy, 2021; Gill, P. 'Terrorist Use of The Internet by the Numbers', Criminology & Public Policy, 2017; Reynolds, S. 'Social Network Analysis of German Foreign Fighters in Syria and Iraq', Terrorism and Political Violence, 2017.

50 For example, the relevance of inadequate mental health provision can be seen in the fact that users looking to join or engage with violent far right organisations have been found to be 31% more likely to engage with mental health. Moonshot, 'Mental Health and Violent Extremism', 2018.

51 See for example 'Digital Citizenship Education: Programming Toolkit', ISD.

52 See for example the following toolkit and resource guide: Tuck, H. 'The Counter-Narrative Handbook', ISD, 2016; GIFCT, Resource Guide.

53 For a useful summary of newer models for measuring behavioural change and sentiment analysis, see: Saltman, E. 'New Models for Deploying Counterspeech: Measuring Behavioral Change and Sentiment Analysis', Studies in Conflict & Terrorism, 2021.

54 Charities Aid Foundation (CAF), 'Machine-Made Goods', 2021.

55 e.g Facebook's oversight board: 'Oversight Board | Independent Judgement. Transparency. Legitimacy.', Oversightboard.com, 2021.

56 Macdonald, S. 'Social Media, Terrorist Content Prohibitions and the Rule of Law', George Washington Program on Extremism, 2019.

57 UNDP 'M&E For Preventing Violent Extremism (PVE)' toolkit, 2021.

58 Kaye, D. Speech Police, Columbia Global Reports, 2019.


UN
DP