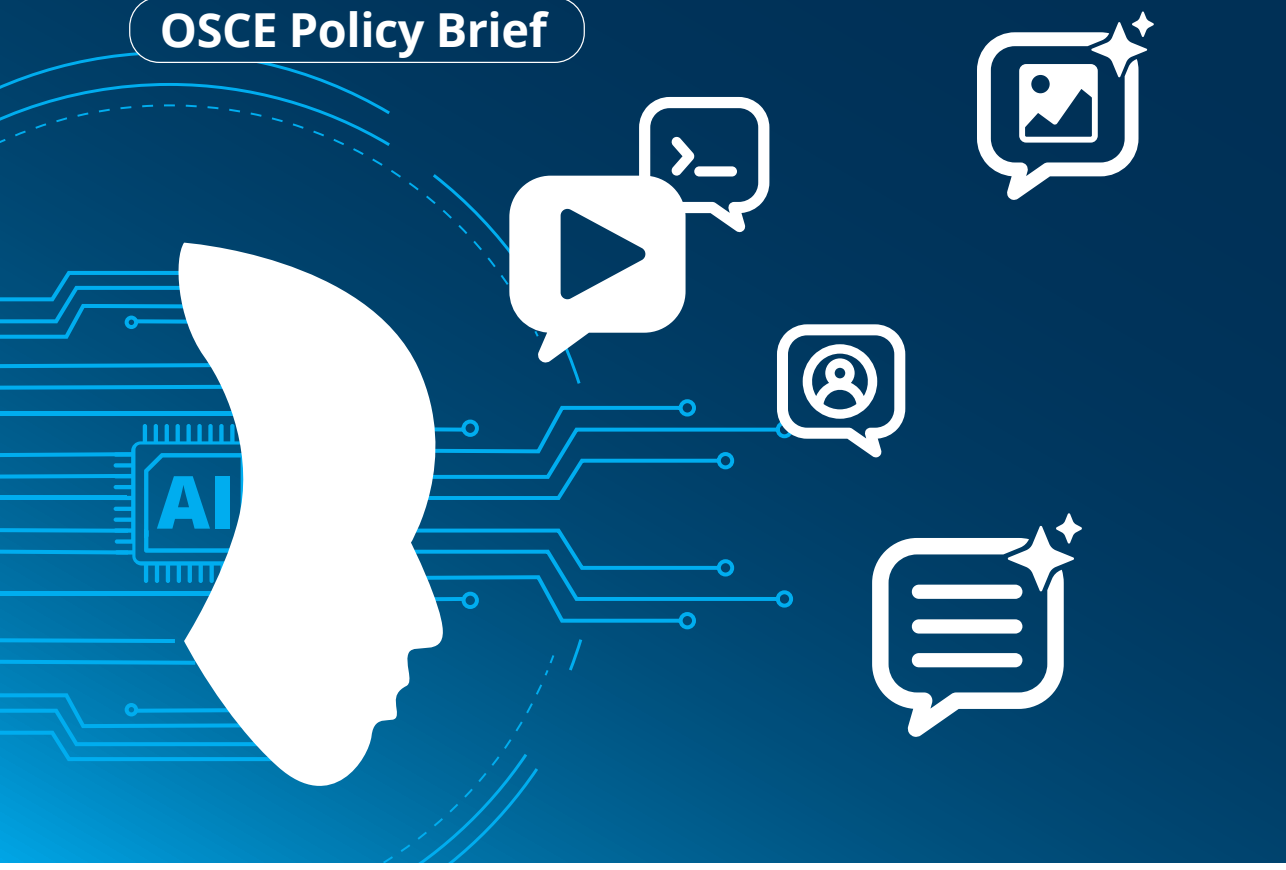
DECEMBER 2025

Ethical Use of Generative Artificial Intelligence in the Context of Preventing and Countering Violent Extremism and Radicalization that Lead to Terrorism

OSCE Policy Brief



OSCE Organization for Security and Co-operation in Europe

Published by the Organization for Security and Co-operation in Europe
Vienna, December 2025

ISBN 978-92-9271-536-6

Transnational Threats Department / Action against Terrorism Unit (TNTD/ATU)
OSCE Secretariat
Wallnerstrasse 6, A-1010 Vienna, Austria
www.osce.org/atu

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This policy brief is intended to provide guidance on the human rights-compliant use of Generative Artificial Intelligence (Gen AI) for policymakers and practitioners in preventing and countering violent extremism and radicalization that lead to terrorism (P/CVERLT) — including government authorities and civil society actors.

There is growing interest in integrating Gen AI within P/CVERLT and a range of ways in which this is currently happening or could happen in the future. Although many of these use cases may have the potential to support P/CVERLT efforts — including by saving resources, enhancing Monitoring and Evaluation (M&E) efforts and improving understanding of the online environment[1] — their effectiveness remains largely unproven. Simultaneously, most Gen AI use cases also pose significant legal and ethical risks to both P/CVERLT actors and the beneficiaries of their activities.

Given this uncertain picture and the risk of P/CVERLT actors rushing to adopt Gen AI before assessing whether the potential benefits can outweigh the associated risks (or having developed policies and procedures to help mitigate them), this policy brief includes a broad set of actions for the consideration of international organizations, including the Organization for Security and Co-operation in Europe (OSCE), as well as government actors, the private sector and civil society, including:

- P/CVERLT actors that plan to or already work with AI technology should conduct consultations with partners and beneficiaries and consider small-scale pilot projects to determine the contexts in which Gen AI tools may be useful and those in which traditional approaches and methods will remain more effective.
- Given the absence of established good practices and guidance regarding the use of Gen AI, P/CVERLT actors should also prioritize the development of internal policies and guidelines for their activities.[2]
- More funding should be allocated by participating States and the private sector to independent research by civil society and academia on the current use of Gen AI within P/CVERLT, including on current practices across the sector and their effectiveness as well as what safeguards are in place to ensure that Gen AI tools are being used legally, ethically and responsibly. This data should support multi-stakeholder

1 For example, a GenAI-powered analytics tool could rapidly summarize and translate violent extremist content from multiple online platforms, helping practitioners to detect emerging narratives or trends that would be hard to discern manually.

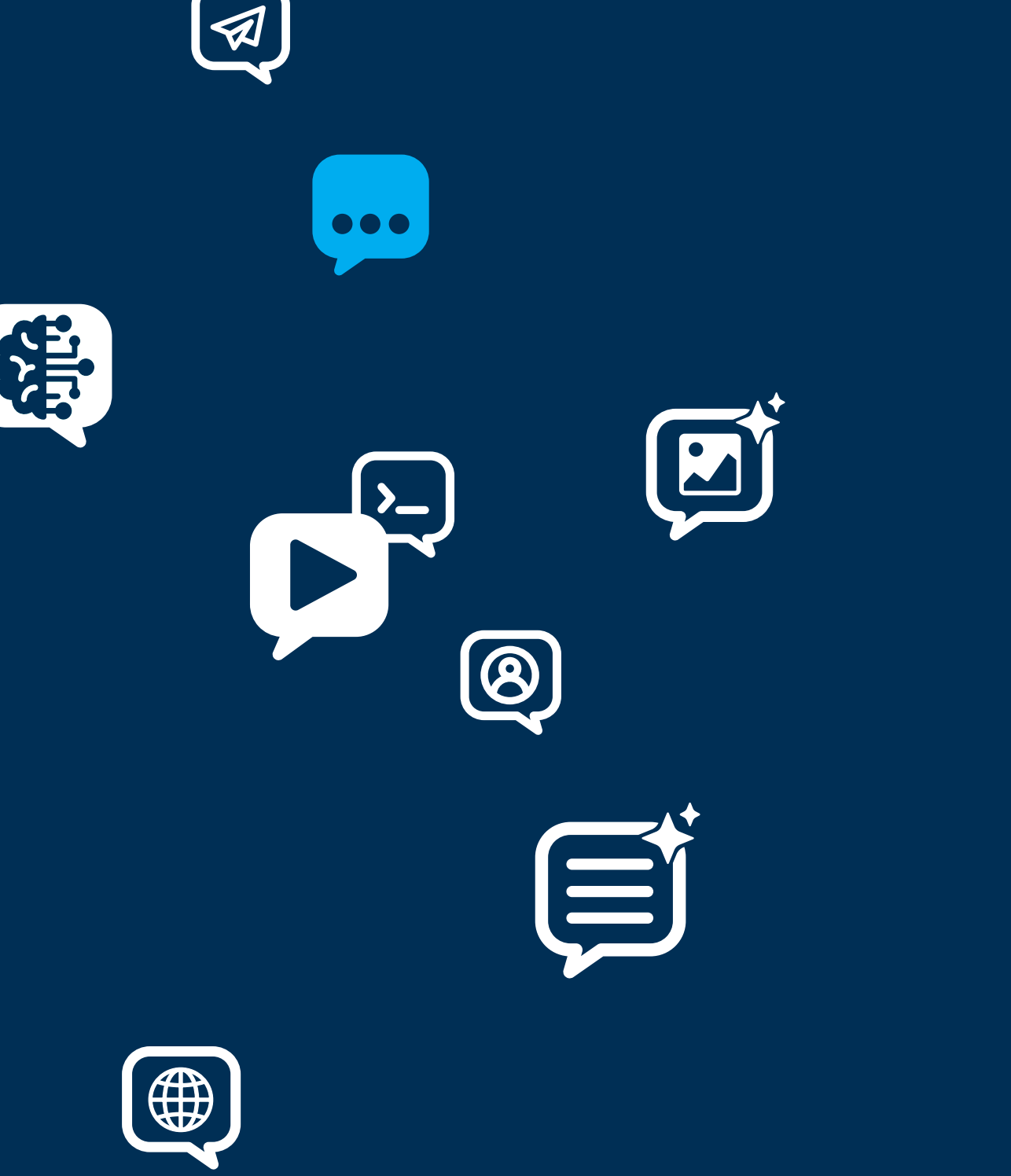
2 For example, United Nations Educational, Scientific and Cultural Organization (UNESCO), "Recommendation on the Ethics of Artificial Intelligence", *UNESCO*, 23 November 2021. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137

efforts to develop good practice or guidance materials on the use of Gen AI in P/CVERLT, grounded in human rights and do-no-harm principles.

- These guidance materials should be operationalized through a range of programmatic activities, including knowledge-sharing and/or awareness-raising and training for P/CVERLT practitioners.
- In addition to the development and operationalization of guidance materials, efforts should be made to develop benchmarks to evaluate Gen AI in the P/CVERLT context.
- The successful use of Gen AI will require different P/CVERLT actors and sectors — notably the private sector — to continue collaborating through multi-stakeholder approaches.
- Considerations of the possible uses of Gen AI should go beyond its potential for saving resources and in-person services and activities.

It is particularly vital to situate Gen AI use in P/CVERLT within the broader context of how such AI systems and models are developed and governed. Issues such as concentrated tech ownership, proprietary training data (often scraped without consent) and market monopolies in AI development raise significant ethical and human rights questions.[3]

3 See, for example, Perrigo, B., "Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic", *Time*, 18 January 2023. Available at: https://time.com/6247678/openai-chatgpt-kenya-workers/



GLOSSARY

This glossary serves to clarify key terms used in this guidebook. The definitions are for this guidebook only and are not official OSCE definitions.

- **AI-generated or AI-enhanced:** Generative AI models and tools are being trained on multi-format datasets using deep-learning techniques, allowing them to respond to prompts (or queries) by generating statistically probable outputs. These outputs can include a wide range of novel content (including video, audio, text and images) – AI-generated – or changes to existing content, AI-enhanced.[4]
- **Chat(ro)bots:** Piece of software or programme designed to conduct simulated conversations with human users. Gen AI-powered chatbots are a significant advance on existing "rules-based" chatbots, with their enhanced predictive capability allowing them to more effectively mimic human speech patterns.[5]
- **Digital literacy:** Ability to define, access, manage, integrate, communicate, evaluate and create information safely and appropriately through digital technologies and networked devices for participation in economic and social life. Digital literacy training on Gen AI could cover how integral bias is within Gen AI tools and services, including how training datasets are selected (and by whom) and the motivations and background of the companies developing Gen AI tools, helping beneficiaries to interpret and assess Gen AI outputs.[6]
- **Guardrails:** Sets of guiding principles and controls developed to govern the outputs of AI systems. They can act as a safeguard against misuse, bias and unethical or illegal practices, but can also refine outputs over time when inaccuracies or other issues are identified within the generated content.[7]
- **Large language models (LLM):** AI programme trained on a large volume of data, using machine learning techniques (deep learning). The size of this training data, and subsequent fine tuning to a particular task, allows the LLM to recognize and interpret inputs from humans and provide a probabilistic output (e.g., the statistically most

4 Definition adapted from Mucci, T., "What is AI-generated content?", *International Business Machines Corporation (IBM)*, 27 November 2024. Available at: https://www.ibm.com/think/insights/ai-generated-content

5 Definition adapted from Robert, J., "Chatbot: definition, uses and impact on companies", *DataScientest*, 4 November 2024. Available at: https://datascientest.com/chatbot-tout-savoir

6 Definition adapted from Organization for Security and Co-operation in Europe (OSCE), "Strengthening Media and Information Literacy in the Context of Preventing Violent Extremism and Radicalization that Lead to Terrorism: A Focus on South-Eastern Europe", *OSCE*, September 2024. Available at: https://www.osce.org/files/f/documents/4/4/575970.pdf

7 Definition adapted from McKinsey & Company, "What are AI guardrails?", *McKinsey & Company*, 14 November 2024. Available at: https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-are-ai-guardrails

likely response). Small language models, on the other hand, are typically trained on a smaller, more focused dataset, and designed with a more specific, tailored objective.[8]

- Open-source AI models: AI model that can be used or modified for any purpose (including to change its outputs) without seeking the permission of its initial developers. There should also be transparency regarding the model's training data and source code, allowing researchers to study its components and how it works.[9]
- Synthetic information/synthetic media: Information that is partially or wholly generated by AI. This can include text, video, imagery and sound. 'Deepfakes' are a specific sub-category of synthetic media.[10]

8 Definition adapted from Cloudfare, "What is a large language model (LLM)?", *Cloudfare*, n.d. Available at: https://www.cloudflare.com/en-gb/learning/ai/what-is-large-language-model/

9 Definition adapted from Williams, R. and O'Donnell, J., "We finally have a definition for open-source AI", MIT Technology Review, 22 August 2024. Available at: https://www.technologyreview.com/2024/08/22/1097224/we-finally-have-a-definition-for-open-source-ai/

10 Definition adapted from "The rise and risks of synthetic media", *Digiwatch* 9 July 2025. Available at: https://dig.watch/updates/the-rise-and-risks-of-synthetic-media

CONTEXT

The OSCE Secretariat's Transnational Threats Department/Action against Terrorism Unit (TNTD/ATU) acts as the focal point and information resource and implementation partner on OSCE's efforts in P/CVERLT and counter-terrorism. This includes multiple initiatives to prevent and counter the misuse of the internet for violent extremist and terrorist purposes while ensuring respect for freedom of expression and other human rights, helping to increase resilience to violent extremist and terrorist content in the online space.

One of these initiatives, project INFORMED (launched in 2023),[11] contributes to closing the gap between research and programming, and support stakeholders from across government and different professional sectors — as well as civil society and communities — to jointly address VERLT in the online space by strengthening media and information literacy (MIL). The project, which focuses on age- and gender-sensitive as well as do-no-harm approaches, includes a broad range of programmatic activities in South-Eastern Europe and Central Asia, which have the objective of delivering locally relevant approaches to tackling violent extremism. Research papers and policy guidance on relevant topics, such as this one, are also produced under the project.

In the framework of its programmatic work, TNTD/ATU has identified significant interest among a variety of stakeholders in the use of Gen AI as a potential tool within P/CVERLT efforts, including in the context of MIL. As a result, TNTD/ATU and the OSCE Office of the Representative on the Freedom of the Media have been co-operating more broadly on the topic, including through organizing a number of joint events, e.g., an expert-level meeting on the use of AI in P/CVERLT together with the International Centre for Counter-Terrorism in March 2024.[12] In December 2024, project INFORMED supported a regional civil society workshop in Tashkent, Uzbekistan (organized by Meta) on the use of Gen AI in P/CVERLT. TNTD/ATU also launched an expert webinar series in July 2024, focused on academic, practitioner and policy approaches to addressing challenges related to the use of emerging technologies for VERLT purposes.[13] The series explores how the human rights-based use of new technologies can be utilized by practitioners in their responses to P/VERLT online.

11 For more information, see OSCE, "INFORMED: Information and Media Literacy in Preventing Violent Extremism – Human rights-based and gender-sensitive approaches to addressing the digital information disorder", *OSCE*, n.d. Available at: https://www.osce.org/project/INFORMED

12 For more information, see OSCE, "Summary Document of Expert-Level Event: Artificial Intelligence in the Context of Preventing and Countering Violent Extremism and Terrorism: Challenges, Risks and Opportunities", *OSCE*, n.d. Available at: https://www.osce.org/files/f/documents/4/f/575877.pdf

13 For more information, see OSCE, "Concept Note 'Facing Division: Preventing and Countering Violent Extremist and Terrorist Content Online: Human Rights-, Age- and Gender-Sensitive Approaches'", *OSCE*, 2024. (Internal document)

Participants and experts in all of the above-mentioned events have called for more guidance from the international community and organizations such as the OSCE on the topic, while underscoring the need for a do-no-harm approach rooted in a robust human rights and ethical framework.[14] This policy brief responds to these requests by aiming to provide policymakers, practitioners and the OSCE with an overview of some of the potential use cases for Gen AI in P/CVERLT and MIL, while considering ethical, legal, and human rights dimensions and risks. It will not address longer-standing uses of other forms of AI in P/CVERLT, including the use of AI-powered tools and instruments by law enforcement and security agencies for counter-terrorism purposes (e.g., machine learning in content moderation, monitoring of social media activities and communications, surveillance, application of Open-Source Intelligence, pattern recognition techniques, deployment of predictive analytics, etc.).

14 For more information, see OSCE, "OSCE enhances media and information literacy skills to effectively prevent and counter violent extremism", *OSCE*, 18 December 2024. Available at: https://www.osce.org/secretariat/583642

INTRODUCTION

UNDERSTANDING GEN AI WITHIN THE BROADER AI LANDSCAPE

Since the release of OpenAI's ChatGPT 3 in November 2022, Gen AI has gone from being a relatively niche technology, primarily of interest to researchers and technologists, to a suite of tools that are being increasingly embedded within a range of businesses and different areas of life (education, communication, etc.). However, it is important to distinguish between "Artificial Intelligence" and "Generative Artificial Intelligence". Although often used interchangeably, Gen AI is one subset of the broader AI field.

AI refers to any computer system or software that performs tasks normally requiring human intelligence.[15] Gen AI, by contrast, refers to machine-learning models designed to generate content (such as text, images, video, audio).[16] Unlike traditional AI systems, which focus on analysing or classifying existing data, Gen AI models learn from existing data patterns to generate contextually relevant content.[17] This relationship is illustrated in the diagram below, (see figure 1) which shows Gen AI as one branch within the overall AI hierarchy.[18]

As a result of the rapid expansion of the capability and accessibility of Gen AI tools, malicious actors — including terrorist and violent extremist actors from across the ideological spectrum — have begun testing different ways in which they can further their aims.

Gen AI's language capabilities have already allowed terrorists and violent extremists to rapidly and economically produce and hyper-focus propaganda material in multiple languages and across multiple formats, including text, images, songs, etc.[19] There have also been relatively unsophisticated efforts to develop, tweak or manipulate AI chatbots for

15 For more information, see Microsoft, "Generative AI vs. other AI types", *Microsoft AI*, n.d. Available at: https://www.microsoft.com/en-us/ai/ai-101/generative-ai-vs-other-types-of-ai#:~:text=,feedback%20loops%2C%20the%20system%20or

16 For more information, see Organisation for Economic Co-operation and Development (OECD), "Generative AI", *OECD*, n.d. Available at: https://www.oecd.org/en/topics/generative-ai.html

17 For more information, see Accenture, "What is generative AI", *Accenture*, n.d. Available at: https://www.accenture.com/us-en/insights/generative-ai

18 Popova Zhuhadar, L., "A Comparative View of AI, Machine Learning, Deep Learning, and Generative AI", *Wikimedia Commons*, 30 July 2023. Available at: https://commons.wikimedia.org/wiki/File:Unraveling_AI_Complexity_-_A_Comparative_View_of_AI,_Machine_Learning,_Deep_Learning,_and_Generative_AI.jpg

19 Humphrys, S., "Analysis: How jihadists experimented with AI in 2024", *BBC Monitoring*, 12 November 2024. Available at: https://monitoring.bbc.co.uk/product/b0002qiw

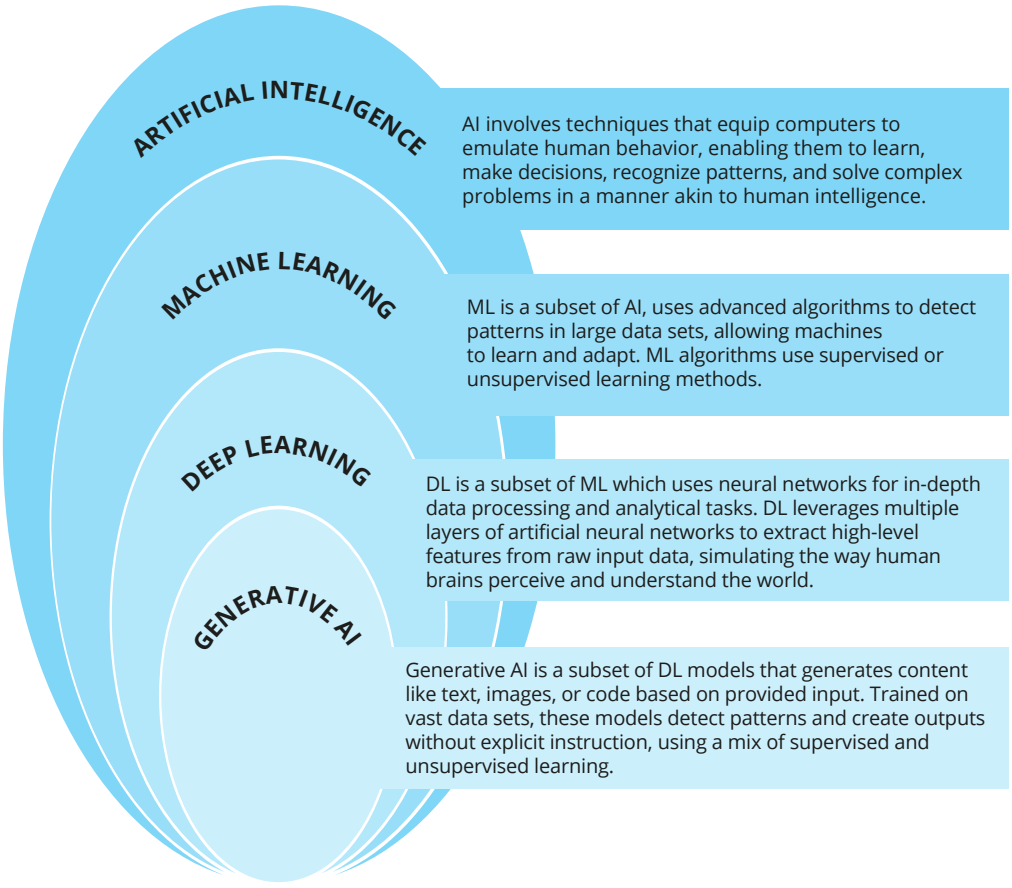


FIGURE 1

Unravelling AI Complexity: A Comparative View of AI, Machine Learning, Deep Learning, and Generative AI (Dr. Lily Popova Zhuhadar, 07.09.2023)

terrorist or violent extremist aims.[20] Both of these use cases provide terrorist and violent extremist actors with the ability to more easily tailor their propaganda and recruitment efforts to specific groups, and to do so at speed and at scale. Although significant, this use case has been relatively limited (to date), with other non-AI tools and techniques continuing to be critical to terrorist exploitation of the internet.

In parallel, other malicious and non-malicious actors have been responsible for the widespread dissemination of a diverse range of AI-generated content online. As a result, the online information environment has become increasingly flooded with AI-generated material, including in disinformation contexts. This trend is significantly outpacing efforts to flag or identify such material as AI-generated (e.g., watermarking[21]), making it even more difficult for individual users to confirm the veracity of the information they interact with.[22] These developments point to the renewed importance and urgency of providing individuals and communities with MIL skills and tools to identify mis-, dis- and mal-information online, including synthetic media

20 Wells, D., "The Next Paradigm Shattering Threat? Right-sizing the Potential Impacts of Generative AI on Terrorism", *Middle East Institute*, 18 March 2024.
Available at: https://www.mei.edu/sites/default/files/2024-03/Wells%20-%20The%20Next%20Paradigm-Shattering%20Threat%20Right-Sizing%20the%20Potential%20Impacts%20of%20Generative%20AI%20on%20Terrorism.pdf;
Siegel, D., "'RedPilled AI': A New Weapon for Online Radicalisation on 4chan", *G-NET Insights*, 7 June 2023.
Available at: https://gnet-research.org/2023/06/07/redpilled-ai-a-new-weapon-for-online-radicalisation-on-4chan/

21 Hulick, K., "Google now adds watermarks to all its AI-generated content", *Science News Explores*, 11 December 2024.
Available at: https://www.snexplores.org/article/google-ai-watermarks

22 Martino, M., "Artificial intelligence is flooding the internet with fake images, video and audio. Can you tell real from fake?", *ABC News*, 13 September 2024.
Available at: https://www.abc.net.au/news/2024-09-14/artificial-intelligence-real-fake-quiz-abc-news-verify/104148236

THE P/CVERLT STATUS QUO

WHAT ARE THE KEY CHALLENGES?

Before seeking to identify and understand the ways in which Gen AI can be used in P/CVERLT, it is important that P/CVERLT actors and policymakers view potential use cases as more than isolated (and often exciting) technological capabilities. Instead, given the risks posed by the use of Gen AI, its adoption should be placed in the broad context of the key current challenges faced by P/CVERLT actors; in other words, which problems can Gen AI safely and ethically help to address, and how severe or persistent are these problems?

Given the scope of this paper, its analysis of the state of the P/CVERLT sector will be brief and focus only on some of the longstanding, well-established challenges. These include:

1. Lack of trust between different sectors: Most national and regional contexts have seen the adoption of whole-of-government and multi-stakeholder approaches to P/CVERLT, in which a variety of local and national government actors work alongside CSOs and other relevant partners, including the private sector. However, as a result of a longstanding lack of trust between government and non-government actors, and differing strategic visions and priorities,[23] these approaches face issues sometimes negatively impacting the effectiveness of the associated P/CVERLT activity.
2. Insufficient funding/resources: Longer-term preventative work typically receives less funding than short-term (often reactive) security-oriented responses. This funding imbalance is exacerbated by other factors, including donors having shorter funding cycles than the optimal lifecycle of a preventative project and/or expectations that may exceed what is realistically achievable or measurable, particularly in the context of P/CVERLT activities online.[24]
3. Difficulties understanding the online landscape: Monitoring and understanding the online ecosystem and patterns of behaviour among key demographics (e.g., youth), and combining these with trends specific to violent extremist activity online, are a critical part of designing and implementing P/CVERLT activities, particularly in the field of MIL. Doing this successfully is difficult — with a diverse range of online platforms of varying sizes exploited by violent extremists and a significant role played by lawful content or activity that can support violent extremist or terrorist narratives (and is

23 Rahlf, L., "Preventing and Countering Violent Extremism in Europe: Expert Views on Contemporary Challenges", *VORTEX*, 3 July 2024. Available at: https://vortex.uni.mau.se/2024/07/preventing-and-countering-violent-extremism-in-europe-expert-views-on-contemporary-challenges/

24 Radicalisation Awareness Network (RAN), "Digital frontrunners: Key challenges and recommendations for online P/CVE work", *RAN Practitioners*, 16-17 June 2022. Available at: https://home-affairs.ec.europa.eu/system/files/2022-08/ran_cn_digital_frontrunners_riga_16-17062022_en.pdf

often amplified by platform algorithms) — and requires expertise and resources typically beyond most CSOs.[25] It can also prove challenging for government representatives.

4. **Understanding what works and what does not:** Perhaps the most persistent P/CVERLT challenge is how to monitor and evaluate its impact and effectiveness. Central to this is the counterfactual nature of prevention — i.e., demonstrating that something has not happened because of the intervention — but there are further challenges relating to data collection, expertise and resources. Evaluations are also often conducted as a prerequisite to securing future funding, calling into question their objectivity.[26]

Although this is not an exhaustive list of the current challenges faced by P/CVERLT actors within the OSCE area, a cursory analysis would suggest that the use of Gen AI might play a role in addressing challenges 2, 3 and 4. The potential lack of trust between different sectors highlights the need to emphasize multi-stakeholder approaches and joint consultations in the introduction of any use cases for Gen AI.

25 Radicalisation Awareness Network (RAN), "Current challenges and solutions related to working with youth on P/CVE", *RAN Practitioners*, 1 December 2022.
Available at: https://home-affairs.ec.europa.eu/document/download/4cee01e1-94d3-46c7-8988-e57b0c2c0e85_en?filename=current_challenges_solutions_related_to_working_youth_on_pcve_22092022_en.pdf

26 Bressan, S., Ebbecke, S., and Rahlf, L., "How Do We Know What Works in Preventing Violent Extremism? Evidence and Trends in Evaluation from 14 Countries", *Global Public Policy Institute*, July 2024.
Available at: https://gppi.net/assets/BressanEbbeckeRahlf_How-Do-We-Know-What-Works-in-Preventing-Violent-Extremism_2024_final.pdf

USING GEN AI IN P/CVERLT

CURRENT OR POTENTIAL APPROACHES

The following section will explore some of the ways in which Gen AI is or could be used in P/CVERLT and will try to ascertain whether these use cases address some of the above-mentioned key challenges. Given how recently Gen AI tools have become widely available and accessible, there are relatively few public examples of its use within P/CVERLT activities. Where these examples exist, the associated projects and programmes have not yet shared monitoring and evaluation data to allow for a meaningful assessment of Gen AI's effectiveness and impact, or information on the steps taken to mitigate the potential risks associated with its use. To address this gap, the section will include approaches known to have been implemented by P/CVERLT actors and approaches that might be broadly feasible using existing or forthcoming Gen AI tools and technologies.

CONTENT CREATION AND TRANSLATION

Firstly, and perhaps most significantly, as with terrorist and violent extremist actors, a range of commercial Gen AI tools offer P/CVERLT actors the opportunity to create multi-format, multi-language and highly tailored content, and to do that more quickly and cheaply than it currently is. This could be based on existing materials — for example, translating them into additional languages, allowing a P/CVERLT organization to engage with previously hard-to-reach groups — or could involve the creation of entirely new ones. Alternatively, a combination of the two could help P/CVERLT actors to brainstorm new approaches and/or receive iterative feedback to assist in the development of new campaigns.

The ethics of using Gen AI to replace specialist services within P/CVERLT will be explored in the subsequent section on Risks and Challenges. However, as already outlined, many P/CVERLT organizations, especially local civil society groups, have limited resources to acquire specialist services, which are often very costly. Gen AI could, therefore, help P/CVERLT actors to use fewer resources to deliver or expand their existing activities at relatively minimal cost to the user, with specialist services (such as translators or videographers) provided for free or via a small subscription fee. Entirely new and highly tailored alternative narratives or strategic messaging campaigns could also be developed significantly more quickly than by employing a consultant or expert, although rigorous human review and oversight would still be required to ensure that the content is accurate and appropriate.

Multiformat content could also be used to make P/CVERLT learning environments or programmatic activity more immersive, helping to tailor the activity to different learning styles and contexts.[27] For example, a December 2024 survey of teenagers in the United States showed that the most widely used social media platforms are video-only,[28] suggesting that MIL programming in that context should prioritize video content over more traditional media formats (which are typically cheaper to access or recreate). Gen AI could enable P/CVERLT actors to easily create a variety of engaging video content for MIL programming, better preparing beneficiaries to navigate the online environment they are experiencing.

Finally, these Gen AI tools could also be used to translate or summarize violent extremist content (depending on tool-specific guardrails and rules), helping P/CVERLT organizations to better understand the online ecosystem they are seeking to counter. Theoretically then, using this type of Gen AI tool could help address two key P/CVERLT challenges (resourcing and understanding the ecosystem).

SUPPORTING MIL IN THE CONTEXT OF P/CVERLT AND TRAINING/EDUCATION

Gen AI tools and techniques are being rapidly rolled out across the education sector, an area that has significant crossover with many P/CVERLT activities, particularly within the MIL context. There are a variety of potential use cases in this context, with Gen AI companies and some educational institutions promoting its use for creating lesson plans, marking assignments and developing curricula.[29] Its proponents argue that rather than replacing educators, Gen AI will free up time for teachers and trainers to more meaningfully engage with their students.[30]

Gen AI tools could allow P/CVERLT and MIL entities to more quickly develop or adapt training and educational materials in response to new and emerging trends, better prioritize how they use limited resources (prioritizing direct engagement with beneficiaries) and address the knowledge gaps of educators and trainers. However, it should be noted that many of the examples of Gen AI use cases in education rely on commercially available

27 See for example, Kenens, A. and Ivanovic, J., "Using generative AI for crisis foresight", *United Nations Development Programme Europe and Central Asia*, 20 November 2024.
Available at: https://www.undp.org/eurasia/blog/using-generative-ai-crisis-foresight

28 Faverio, M. and Sidoti, O., "Teens, Social Media and Technology", *Pew Research Center*, 12 December 2024.
Available at: https://www.pewresearch.org/internet/2024/12/12/teens-social-media-and-technology-2024/

29 Brenner, S., "Comparative lit class will be first in Humanities Division to use UCLA-developed AI system", *UCLA Newsroom*, 4 December 2024.
Available at: https://newsroom.ucla.edu/stories/comparative-literature-zrinka-stahuljak-artificial-intelligence?taid=6755debeb8836e000133c355&utm_campaign=trueAnthem_manual&utm_medium=trueAnthem&utm_source=twitter

30 Agarwal, A., "How AI Can Revive a Love of Learning", *New York Times*, 7 December 2024.
Available at: https://www.nytimes.com/2024/12/07/special-series/artificial-intelligence-schools-education.html

(and potentially costly) tools, in-house tools developed at significant cost by large institutions or chatbots that are currently free, but where transparency regarding the data they are trained on is lacking.

A wider integration of Gen AI within MIL and P/CVERLT educational initiatives — both within learning experience and curricula — could also assist beneficiaries in developing a greater understanding of working with AI, including by building transversal skills (such as critical thinking) that can help them navigate an increasingly AI-infused online and workplace environment.[31] However, this will require a careful balancing act, as one January 2025 study found a significant negative correlation between frequent use of AI tools and critical thinking abilities, emphasizing the need for educational strategies that promote critical engagement with AI technologies.[32] Teaching beneficiaries how to identify and respond to AI-generated material online is also quickly becoming critical to MIL initiatives. There is, therefore, a clear significant role for technology (including Gen AI) in training beneficiaries, such as through partnerships with private sector companies, some of which are themselves responsible for enabling the creation and publication of inauthentic material.

Finally, Gen AI tools (e.g., learning assistants) could help P/CVERLT providers to better monitor the effectiveness and impacts of their programmes, assisting their M&E efforts. For example, a Gen AI learning assistant could seamlessly collect existing (or additional) metrics before and during MIL training, and contribute towards its sustainability by providing ongoing support and advice after any formal training. This type of use case could then assist the P/CVERLT sector in addressing two of their key challenges (resources and M&E).

THE USE OF CHATBOTS IN P/CVERLT

Beyond these broad, widely applicable use cases, there are more specific areas in which Gen AI could assist (and is already assisting) P/CVERLT practitioners, particularly through the use of chatbots. It should be noted, however, that there are significant risks associated with the use of chatbots in the context of P/CVERLT — including in relation to data privacy, bias and transparency — which will be explored in greater detail in the below Risks and Challenges section. The subsequent list of use cases should, therefore, be understood as an overview of already existing practices, rather than a recommendation for programming.

31 Crosling, G., Atherton, G. and Azizan, S.N., "The importance of TVET students' critical and flexible thinking skills for AI competence", *United Nations Educational, Scientific and Cultural Organization (UNESCO)*, October 2023. Available at: https://unevoc.unesco.org/up/TVET_students_AI_competence.pdf

32 Gerlich, M., "AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking", *Centre for Strategic Corporate Foresight and Sustainability*, Swiss Business School, 3 January 2025. Available at: https://www.mdpi.com/2075-4698/15/1/6

A September 2024 study provides a helpful use case, indicating that chatbots can be effective in countering conspiracy theories, which sometimes constitute pathways to violent extremism and terrorism. US researchers developed "DebunkBot", which appears to have a durable impact on users' belief in a variety of conspiracy theories; interestingly, a follow-up study found that a purely factual approach was more effective than one which was persuasive but did not present factual evidence.[33] There are a variety of ways in which this type of approach could be integrated into MIL activities, with the potential to address specific narratives or ideologies, or to counter polarizing or toxic behaviour online.

Projects are already underway to deploy chatbots in significantly higher risk contexts, including using a chatbot trained on a database of insights from P/CVERLT and disengagement programming to 'de-radicalize' individuals involved in respective rehabilitation and reintegration programmes, and perform individual automated risk assessments.[34] This type of approach could in theory be used to deliver personalized counter-narratives, which could be tested and iterated over time, offering 24/7 tailored support in multiple languages, in the tone and style of at-risk individuals or groups.[35]

Chatbots could also be developed to provide support to P/CVERLT practitioners and policymakers. For example, a chatbot interface could be trained on a range of terrorist or violent extremist data (e.g., a dataset of current online violent extremist activity or of the latest research into terrorism and violent extremism), allowing policymakers and practitioners to more easily understand and engage with data critical to their work. Otherwise, examples of testing the effectiveness of alternative narrative or disengagement techniques on 'terrorist chatbots' have been used in attempts to give insights on the mindset of radicalized individuals.[36]

It is worth emphasizing, however, that, although steps could be taken to mitigate risk and bias in these scenarios, significant challenges would remain. Further, few of these examples are 'off-the-shelf' capabilities or could be easily developed using commercially available Gen AI software. Most would require access to a significant volume of terrorist and violent extremist training data, as well as to specific technology and technological skills, placing them beyond the reach of most CSOs. As such, their implementation is likely to

33 Walsh, D., "MIT Study: An AI chatbot can reduce belief in conspiracy theories", *Massachusetts Institute of Technology (MIT) Management Sloan School*, 30 September 2024.
Available at: https://mitsloan.mit.edu/ideas-made-to-matter/mit-study-ai-chatbot-can-reduce-belief-conspiracy-theories#:~:text=The%20artificial%20intelligence%2Dpowered%20%E2%80%9CDebunkBot,according%20to%20a%20new%20study

34 Information provided to the author through their engagement with P/CVERLT actors.

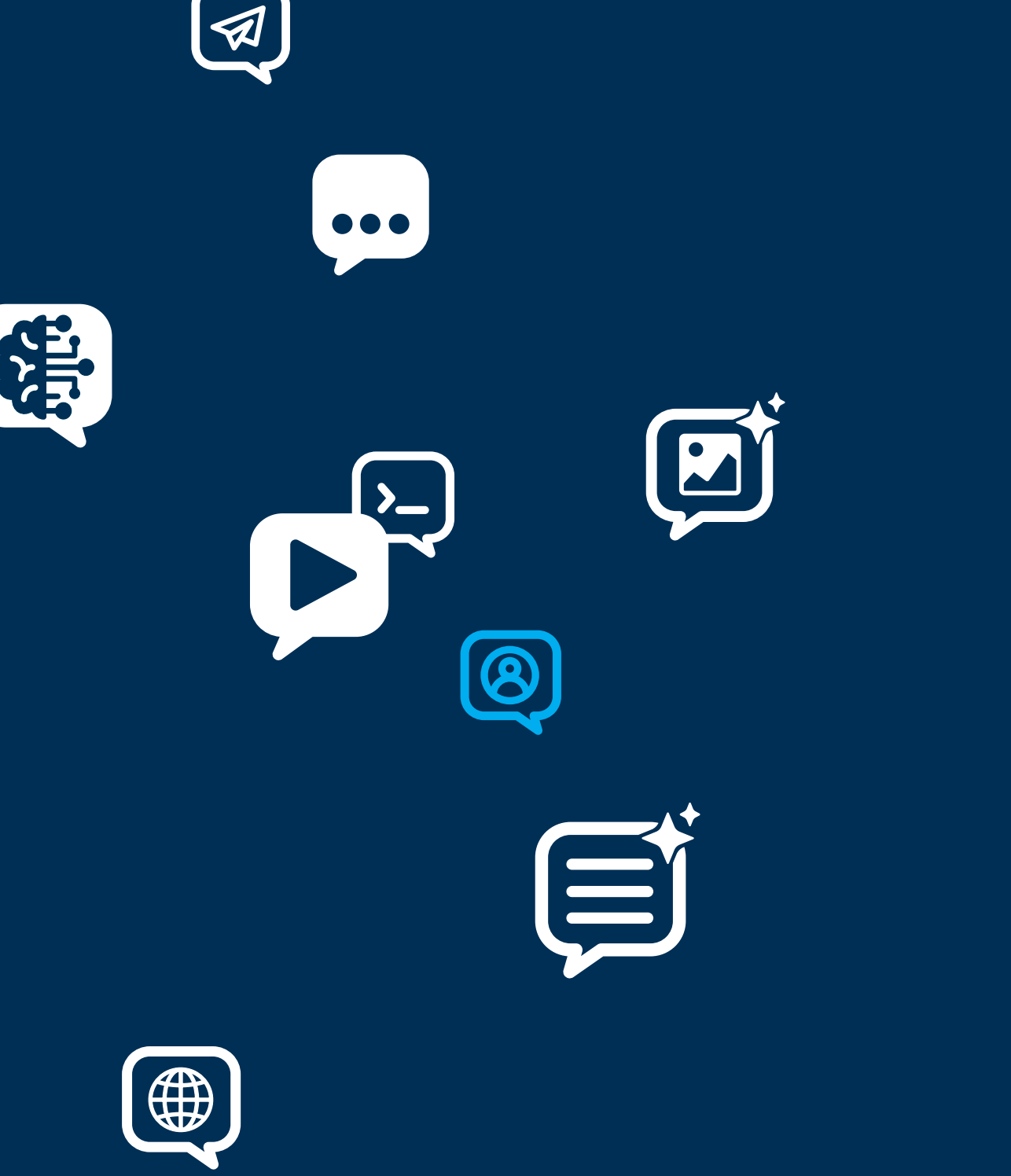
35 Peredaryenko, M. and Heng, A., "Beyond bullets: Leveraging AI and VR in P/CVE efforts", *Awani International*, 13 December 2024. Available at: https://international.astroawani.com/malaysia-news/columnist-beyond-bullets-leveraging-ai-and-vr-pcve-efforts-500455

36 Mathur, P., Broekaert, C. and Clarke, C.P., "The Radicalization (and Counter-radicalization) Potential of Artificial Intelligence", *International Centre for Counter-Terrorism (ICCT)*, 1 May 2024.
Available at: https://icct.nl/publication/radicalization-and-counter-radicalization-potential-artificial-intelligence

require a multi-stakeholder approach that includes government, civil society and the private sector.

In this context, however, it is important to highlight that to strengthen accountability and ensure responsible deployment, the benchmarking of AI systems is essential. AI benchmarks are standardized evaluation tools used across sectors — from healthcare and law to education[37] — to assess whether AI systems perform safely, reliably and fairly before their deployment. They help prevent harmful errors by offering consistent ways to test an AI model's capabilities against real-world scenarios. The development of benchmarks for Gen AI in P/CVERLT is still at an early stage, with no consensus standards in place. Without them, P/CVERLT actors risk deploying AI tools that misidentify threats, reproduce bias or violate rights, especially when distinguishing, for instance, between non-violent extremism and incitement to violence. Grounded in international law, human rights and established good practices in P/CVERLT policy support and programmatic work, international organizations such as the OSCE are well positioned to lead the development of specialized benchmarks. This would provide P/CVERLT actors with clear standards to ensure that emerging tools meet ethical, human rights-compliant and operational expectations. The development of such benchmarks, anchored in international human rights standards, is an urgent next step to maximize AI's benefits for security while minimizing the risk that it will do more harm than good when used in a P/CVERLT context.

37 Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepano, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J. and Tseng, V., "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models", *PLOS Digital Health*, 9 February 2023.
Available at: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000198#:~:text=In%20this%20study%2C%20we%20evaluate,established%2C%20showing



USING GEN AI IN P/CVERLT

RISKS AND CHALLENGES

To properly contextualize the risks associated with the above-mentioned potential or actual use cases, it is perhaps helpful to view them on a spectrum. At one of its ends is the use of Gen AI in day-to-day operational or organizational activities such as summarizing or translating documents or drafting emails. In this context, Gen AI has relatively small but tangible benefits (saving time and resources), but humans remain in the loop and broadly capable of conducting meaningful oversight to identify issues such as inaccuracy. As a result, these activities have a typically low (but not insignificant) risk profile, although these risks may increase over time if trust in the technology increases and oversight decreases.

At the other end of the spectrum are more complex activities such as the development and deployment of P/CVERLT chatbots. Although Gen AI has the potential to deliver much greater benefits in this context, these benefits have not been realized yet, in part, due to the recency of the initiatives in this space. Correspondingly, however, P/CVERLT actors can to a much lesser extent conduct meaningful oversight of almost every part of the process, including the extent of bias within training data, what data is selected and extracted by the chatbot, and the content of individual outputs. Each of these steps requires high levels of digital literacy, including on AI bias and how to interpret AI output, alongside transparency on the part of the chatbot developer. As a result, this type of activity has a much higher risk profile.

LEGAL/ETHICAL RISKS

INTELLECTUAL PROPERTY

As previously mentioned, there are significant potential legal and ethical risks posed by the use of Gen AI technologies in P/CVERLT. Firstly, most (if not all) of the large Gen AI products have been trained on vast swathes of data scraped from the internet, including copyrighted material.[38] Although there has been a move towards licensing agreements in some jurisdictions, there are ongoing copyright and intellectual property (IP) legal cases against Gen AI companies globally,[39] with media outlets, artists and researchers, among other content creators, arguing that they did not give permission for their IP to be

38 Milmo, D., "'Impossible' to create AI tools like ChatGPT without copyrighted material, OpenAI says", *The Guardian*, 8 January 2024. Available at: https://www.theguardian.com/technology/2024/jan/08/ai-tools-chatgpt-copyrighted-material-openai?CMP=Share_AndroidApp_Other

39 For a sample, see the intellectual property cases and policy tracker on law firm Mishcon de Reya's website. Available at: https://www.mishcon.com/generative-ai-intellectual-property-cases-and-policy-tracker

used in this way or that they did not agree to or get a fair compensation for that use. P/CVERLT organizations may be able to save money by using Gen AI in their activities, but it could come at the expense of individuals who will not have been paid for the re-purposing of their work.

BIAS

Research into the use of AI in a range of contexts, including within law enforcement and other government functions,[40] has demonstrated the implicit and explicit bias that can be caused by both training data sets and the algorithms used to query them. Many of these issues appear to be replicated and exacerbated at scale across Gen AI tools, with clear evidence that Gen AI reflects bias on the basis of gender,[41] sexuality, ethnicity, race and age across a range of contexts including health,[42] the media[43] and law enforcement.[44] This is an issue even in lower risk Gen AI use cases such as summarizing or analysing documents. P/CVERLT users of Gen AI should, therefore, be aware of the bias within its outputs — and approach them with a degree of scepticism and an analytical mindset — and also be wary of how these outputs could perpetuate the existing biases of end users.

To mitigate these biases and/or develop a tool more tailored to their needs, many organizations are customizing open-source models, adding a significant portion of their own internal data to the training dataset. In the P/CVERLT context, this data could include materials from previous campaigns or programmes, good practices and research insights. Although this approach allows for greater control over the tool's outputs, P/CVERLT actors may still have problems understanding what training data has been used in creating the baseline model and, therefore, what biases might be inherent to it.

40 For a specific sample see: Ungoed-Thomas, J. and Abdulahi, Y., "Warnings AI tools used by government on UK public are 'racist and biased'", *The Guardian*, 25 August 2024.
Available at: https://www.theguardian.com/technology/article/2024/aug/25/register-aims-to-quash-fears-over-racist-and-biased-ai-tools-used-on-uk-public#:~:text=3%20months%20old-,Warnings%20AI%20tools%20used%20by%20government,public%20are%20'racist%20and%20biased'&text=Artificial%20intelligence%20and%20algorithmic%20tools,%E2%80%9Centrenched%E2%80%9D%20racism%20and%20bias

41 United Nations Educational, Scientific and Cultural Organization (UNESCO), "Challenging systematic prejudices: an investigation into bias against women and girls in large language models", *UNESCO*, 7 March 2024.
Available at: https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes

42 Ayoub, N.F., Balakrishnan, K., Ayoub, M.S., Barrett, T.F., David, A.P. and Gray, S.T., "Inherent Bias in Large Language Models: A Random Sampling Analysis", *Mayo Clinical Proceedings Digital Health*, June 2024.
Available at: https://www.mcpdigitalhealth.org/article/S2949-7612(24)00020-8/fulltext

43 See for example, Fang, X., Shangkun, C., Mao, M., Zhang, H., Zhao, M. and Zhao, X., "Bias of AI-generated content: an examination of news produced by large language models", *Nature*, 4 March 2024.
Available at: https://www.nature.com/articles/s41598-024-55686-2

44 Stanley, J., "AI Generated Police Reports Raise Concerns Around Transparency, Bias", *American Civil Liberties Union*, 10 December 2024. Available at: https://www.aclu.org/news/privacy-technology/ai-generated-police-reports-raise-concerns-around-transparency-bias

TRANSPARENCY AND EXPLICABILITY

The need for transparency is critical across all elements of the Gen AI process. This includes what training data has been used and how the tool has been trained on this data, the need to disclose that individuals are interacting with a Gen AI tool, where the end user's data — including their interaction with the tool — is stored, and what policies are in place with regard to sharing it with government authorities.

Transparency is not an end in and of itself, and should always be linked to accountability. The information provided should, therefore, be understandable to the P/CVERLT actor and any end users/beneficiaries, but also provide practical insights that can help them identify ways to mitigate risks. For example, information provided by a commercial vendor regarding the data collection and training phase needs to be made both explicable and understandable for a non-technologist audience. Otherwise, P/CVERLT practitioners risk being unable to explain to partners and beneficiaries how a tool has been created or how it operates, or address understandable concerns regarding bias, discrimination or potential disclosure to authorities. Given the resource challenges many P/CVERLT actors face (particularly CSOs), it seems unlikely that many will be able to develop the specialist knowledge required to achieve this.

USE OF TERRORIST MATERIAL AND LEGAL ISSUES

Some researchers and P/CVERLT practitioners have suggested that Gen AI tools could be trained on data that includes terrorist and violent extremist content. This raises significant ethical and legal issues — will P/CVERLT organizations be legally permitted to collect and use this type of illegal content? If this activity was permitted under national or regional legislation and a chatbot was developed, what safeguards would be in place to limit and monitor its activities? Who would take legal responsibility if the chatbot behaved in unexpected or potentially illegal ways? The last two questions are not hypothetical, with research showing that Gen AI chatbots can be easily 'jailbroken,' allowing users to ask them any question,[45] while chatbot companies have been sued after users of their services died by suicide.[46]

There are no straightforward answers to these questions, given the novelty of these technologies and the relatively sluggish legislative response in most jurisdictions. The EU AI Act represents a positive exception, though it has also faced criticism for containing too

45 Maxwell, T., "AI Chatbots Can Be Jailbroken to Answer Any Question Using Very Simple Loopholes", *Gizmodo*, 20 December 2024. Available at: https://gizmodo.com/ai-chatbots-can-be-jailbroken-to-answer-any-question-using-very-simple-loopholes-2000541157

46 See for example, Pierson, B., "Mother sues AI chatbot company Character.AI, Google over son's suicide", *Reuters*, 24 October 2024. Available at: https://www.reuters.com/legal/mother-sues-ai-chatbot-company-characterai-google-sued-over-sons-suicide-2024-10-23/

many loopholes.[47] Outside these specific use cases (and the IP issues already highlighted), there are a range of other legal risks associated with the use of Gen AI tools in P/CVERLT, including to what extent Gen AI tools comply with data protection and privacy regulation, and whether their use by non-governmental actors is permitted under AI legislation.[48]

CLIMATE IMPACT

The use of Gen generative AI carries significant environmental and climate costs. Although its precise impact is hard to calculate — in large part due to obfuscation by the tech sector[49] — training and operating large language models (LLM) require enormous amounts of energy and huge volumes of water to cool the associated data centres. Although this macro challenge is only indirectly linked to P/CVERLT actors, choosing to actively use Gen AI when its development and use have direct and harmful impacts on the climate crisis does pose serious ethical questions.

EFFECTIVENESS

In addition to the legal and ethical risks associated with using Gen AI in P/CVERLT addressed above, there is a broad range of risks or challenges that might impact the effectiveness of these tools.

INACCURACIES

Gen AI tools continue to be affected by inaccuracies, referred to by the industry as 'hallucinations.'[50] Depending on the P/CVERLT Gen AI use case and the extent to which practitioners are familiar with the context they are operating in, purely factual inaccuracies may be identifiable and rectifiable. However, the more an organization uses Gen AI, the bigger the risk that inaccuracies will compound, with relatively minor inaccuracies becoming a much bigger problem with negative impacts on the performance and reputation of a P/CVERLT actor.

47 European Civic Forum: "Packed with loopholes: Why the AI Act fails to protect civic space and the rule of law", 4 April 2024. Available at: https://civic-forum.eu/advocacy/artificial-intelligence/packed-with-loopholes-why-the-ai-act-fails-to-protect-civic-space-and-the-rule-of-law

48 Interview with Elliot Grainer, author of "AI Ethics and PCVE", *Radicalisation Awareness Network (RAN)*, July 2024.

49 Taylor, C., "How much is AI hurting the planet? Big tech won't tell us", *Mashable*, 3 September 2024. Available at: https://mashable.com/article/ai-environment-energy

50 Kelsey-Sugg, A. and Carrick, D., "AI hallucinations caused artificial intelligence to falsely describe these people as criminals", *ABC News*, 3 November 2024. Available at: https://www.abc.net.au/news/2024-11-04/ai-artificial-intelligence-hallucinations-defamation-chatgpt/104518612

AI-generated content is also polluting the broader information environment, including key sources for the P/CVERLT community such as academia.[51] Any P/CVERLT actor seeking to use evidence-based research as part of its training data for a small or large language model may inadvertently include inaccurate and inauthentic information at its foundational level. This example is symptomatic of a broader issue with building or tweaking AI tools and including publicly accessible information in its training dataset — an increasing volume of online information is AI-generated.

Another area where accuracy is especially important is language, particularly for P/CVERLT activities in regions that have less widely-spoken languages. Most large commercial Gen AI tools have a significant bias[52] towards a relatively small subset of languages, because their training datasets have heavily relied on scraping data from the public web; more than half of all websites are in English, while 90% of websites are written in just ten languages.[53] This is further perpetuated by the widespread dissemination of AI-generated content in these ten languages. This systemic bias[54] means that widely available Gen AI commercial tools are likely to face significantly more accuracy issues with less widely spoken and written languages.

To address these challenges, some countries with less widely spoken languages have begun developing their own LLMs adapted to local linguistic needs[55]. The lessons learned underline the importance of ensuring that such models reflect not only linguistic but also cultural nuances, idiomatic expressions and specific communication practices.[56] An inclusive approach is, therefore, critical — one that involves collaboration across diverse stakeholders — to ensure these tools are developed responsibly and can serve all communities

51 For example, a September 2024 study found 139 papers on Google Scholar that appeared to be AI-generated. For more information, see Haider, J., Soderstrom, K.R., Ekstrom, B. and Rodl, M., "GPT-fabricated scientific papers on Google Scholar: Key features, spread, and implications for pre-empting evidence manipulation", *Harvard Kennedy School Misinformation Review*, 3 September 2024. Availableat:https://misinforeview.hks.harvard.edu/article/gpt-fabricated-scientific-papers-on-google-scholar-key-features-spread-and-implications-for-preempting-evidence-manipulation/

52 For example, ChatGPT-4 recognises sentences in Hausa, Nigeria's most widely spoken language, only 10-20% of the time. Referenced in Moorosi, N., "Better data sets won't solve the problem – we need AI for Africa to be developed in Africa", *Nature*, 11 December 2024. Available at: https://www.nature.com/articles/d41586-024-03988-w

53 These languages are English, Russian, Spanish, German, French, Japanese, Turkish, Portuguese, Italian and Farsi. Cited in Wong, M., "The AI Revolution Is Crushing Thousands of Languages", *The Atlantic*, 12 April 2024. Available at: https://www.theatlantic.com/technology/archive/2024/04/generative-ai-low-resource-languages/678042/

54 Gorbacheva, A., "No language left behind: How to bridge the rapidly evolving AI language gap", *United Nations Development Programme (UNDP) Kazakhstan*, 4 October 2023. Available at: https://innovation.eurasia.undp.org/ai-language-gap/

55 Bajwa, A., "Nations building their own AI models add to Nvidia's growing chip demand", *Reuters*, 29 August 2024. Available at: https://www.reuters.com/technology/nations-building-their-own-ai-models-add-nvidias-growing-chip-demand-2024-08-29/

56 Tao. Y., Viberg, O., Baker, R.S. and Kizilcec, R.F., "Cultural bias and cultural alignment of large language models", *PNAS Nexus* 3(9), 17 September 2024. Available at: https://academic.oup.com/pnasnexus/article/3/9/pgae346/7756548

effectively.[57] This includes meaningful engagement with local stakeholders working in the field of P/CVERLT, including civil society, academia and private sector partners.

CONTEXT AND AUTHENTICITY

Even when Gen AI (or other AI tools such as machine learning) can accurately translate language, important cultural and legal context is often missed.[58] Cultural understanding is particularly critical in the development of P/CVERLT chatbots, as research has consistently emphasized the importance of having an authentic and credible 'messenger' (with victims[59] and 'formers'[60] both cited as examples of authenticity). Although Gen AI can theoretically mimic the views and voice of this type of messenger, gaps and biases in training data may mean that cultural context is absent or misrepresented. There are also broader questions around how 'authentic' any chatbot explicitly labelled as Gen AI may or may not feel to different target audiences.

Emerging research into the impact of Gen AI on language has also suggested that it can result in a text that lacks "the distinctive suite of words and grammar that gives personality to our sentences."[61] This has led experts to speculate that, over time, people might react negatively to this absence of nuance and the inauthenticity of AI-assisted communication. Conversely, the value and impact of truly authentic communications in P/CVERLT may increase.

Finally, this inability to understand nuances or local contexts may create additional legal and ethical risks, depending on the national policy framework in which Gen AI tools are deployed. For example, the use of a chatbot in P/CVERLT contexts, e.g., for disengagement purposes, could identify a disproportionately large number of individuals — or individuals from a particular minority group — as "at risk of radicalization", due to bias or data gaps. This could be particularly problematic in countries with overly broad definitions of extremism given the potential legal ramifications for these individuals of being identified

57 Kannan. P., "Improving Equity and Access to Non-English Large Language Models", *Stanford University Human-Centered Artificial Intelligence*, 22 April 2024.
Available at: https://hai.stanford.edu/news/improving-equity-and-access-non-english-large-language-models

58 Macdonald, S., Mattheis, A. and Wells. D., "Using Artificial Intelligence and Machine Learning to Identify Terrorist Content Online", *Tech Against Terrorism Europe*, 17 January 2024.
Available at: https://static1.squarespace.com/static/63e0c75f41ff767f07530a6f/t/65a5349648ce18428d686126/1705325720146/TATE+-+AI+REPORT+FINAL+%281%29.pdf

59 See for example, Milosevic, A., "Testimonials: Involving victims/survivors of terrorism in P/CVE", *Radicalisation Awareness Network (RAN)*, 22 September 2023. Available at: https://home-affairs.ec.europa.eu/document/download/b0c7b414-6837-4fa2-8599-7a6ad2ca056f_en?filename=ran_testimonials_involving_vot_in_pcve_092023_en.pdf

60 See for example, Wichmann, F., "Harnessing the Power of Victims, Survivors, and Former Extremists in Prevention Campaigns", *Journal EXIT-Deutschland*, 4 September 2023. Available at: https://journal-exit.de/inclusive-voices-unite/

61 Professor Tony Thorne, quoted in Booth, R., "Losing our voice? Fears AI tone-shifting tech could flatten communication", *The Guardian*, 11 December 2024. Available at: https://www.theguardian.com/society/2024/dec/11/ai-tone-shifting-tech-could-flatten-communication-apple-intelligence

as 'extremist', and would pose much broader questions for the P/CVERLT actor regarding privacy, consent and liability.

RESOURCES

There are also potential financial barriers to the effectiveness of Gen AI approaches in P/CVERLT. As previously mentioned, many of the use cases (e.g., chatbots) require significant resources to reach the point of deployment, given the type of data collection, cleaning, testing and tweaking needed to overcome some of the challenges highlighted above. Once deployed, these types of approaches would also require ongoing resources to monitor their activities, alongside technological capabilities to make adjustments as issues or errors are identified. Finally, there is a risk that this could divert funding from what are often already scarce resources for disengagement programmes.

Even in less advanced use cases, there are financial and dependency-related risks. Although P/CVERLT organizations are currently able to utilize a variety of Gen AI services for free (within certain limits), some experts warn that the overall trajectory of the Gen AI sector is downwards,[62] leading to increased monetization of existing products.[63] This raises the risk of P/CVERLT organizations developing dependencies on Gen AI technology within a current, largely free or cheap business model; if this model subsequently shifts, P/CVERLT organizations may be forced to pay significantly increased subscription fees to access their intellectual property or to continue business-as-usual services.

M&E GAPS AND CHALLENGES

Finally, as previously mentioned, longstanding M&E challenges mean there remains a degree of uncertainty regarding what does and does not work in P/CVERLT. There is a risk that building new Gen AI P/CVERLT models based solely on existing models might perpetuate current faulty thinking or reasoning, crystallizing P/CVERLT models and methods, and limiting opportunities for future iteration. Regardless of the Gen AI approaches implemented (and the data behind them), ongoing M&E will be critical to assess their effectiveness and identify any negative impacts.

62 Gray Widder, D. and Hicks, M., "Watching the Generative AI Hype Bubble Deflate", *Harvard Kennedy School*, 20 November 2024. Available at: https://ash.harvard.edu/resources/watching-the-generative-ai-hype-bubble-deflate/

63 Robison, K., "OpenAI is charging $200 a month for an exclusive version of its o1 'reasoning' model", *The Verge*, 5 December 2024. Available at: https://www.theverge.com/2024/12/5/24314147/openai-reasoning-model-o1-strawberry-chatgpt-pro-new-tier

CONCLUSIONS AND RECOMMENDATIONS

A recurring criticism within counter-terrorism and P/CVERLT circles is how slowly the sector responds to the misuse of new technology by terrorists and violent extremists. Simultaneously, the sector has sometimes been too quick to integrate these technologies within their responses and has done so without sufficient consideration of human rights and the management of a range of related risks. Given this context — and following several years of hype regarding the possibilities offered by Gen AI — it is unsurprising that P/CVERLT actors are exploring how this technology might help them address a variety of longstanding issues, including funding pressures, a complex and fast-changing online environment and M&E challenges.

However, it remains uncertain whether any potential benefits offered by Gen AI in P/CVERLT can outweigh a broad range of associated operational, ethical and legal risks. It is, therefore, important that P/CVERLT actors do not rush — or are not rushed by donors and partners — into adopting Gen AI without first considering how and where it will add value to their work, and how to mitigate the very real risks it poses. Until they can mitigate these risks and challenges to an acceptable level, P/CVERLT actors should be cautioned against using Gen AI for P/CVERLT activities.

This risk assessment and mitigation is a significant task for individual P/CVERLT actors to undertake, particularly civil society actors who are struggling with limited resources and a challenging donor environment. To assist them — and the rest of P/CVERLT community — in this process, there is a broad set of actions that the OSCE and other international, regional and national entities, as well as private sector and civil society actors should consider.

RECOMMENDATIONS FOR ALL P/CVERLT ACTORS, REGARDLESS OF SECTOR OR ORGANIZATION

- Gen AI should not primarily be viewed by donors or national authorities as a way to save resources or replace critical in-person services and activities. Supporting the ongoing P/CVERLT efforts of CSOs and other non-government actors should continue to be seen as a much-needed investment in a multi-stakeholder approach.
- The successful use of Gen AI will require different P/CVERLT actors and sectors — notably the private sector — to continue collaborating through inclusive and multi-stakeholder approaches, including by sharing information on emerging good practices, particularly given the challenging resource environment and the speed with which Gen AI technology is developing. These approaches can help to build trust between different P/CVERLT actors and ensure inclusive ways to integrate Gen AI.
- P/CVERLT actors that plan to, or already, work with AI technology should conduct consultations with partners and beneficiaries, and consider small-scale pilot projects to determine the contexts in which Gen AI tools may be useful, and those in which traditional approaches and methods will remain more effective.
- Given the absence of established good practices and guidance regarding the use of Gen AI, P/CVERLT actors should prioritize developing clear internal policies and guidelines for their activities. These policies should be based on human rights and do-no-harm principles, informed by broader guidance on the responsible use of AI. and include an assessment of the risks associated with using Gen AI. They should also include stringent M&E requirements for any integration of Gen AI tools and techniques.
- Good practice or guidance material on the use of Gen AI in P/CVERLT should be developed in consultation with relevant national, regional and international P/CVERLT entities. This guidance should be based on a contemporary understanding of the P/CVERLT landscape and a comprehensive mapping exercise (informed by the above research) and be grounded in human rights and do-no-harm principles.
- This guidance material should be operationalized through a range of programmatic activities, including knowledge-sharing and/or awareness-raising and training for P/CVERLT practitioners. This should include rights-based digital literacy training for policymakers and practitioners, helping them to critically assess the usefulness of any Gen AI tools/techniques that they plan to procure or use, identify potential risks and seek to mitigate these risks through appropriate safeguards or refrain from using them if the risk is found too high and cannot be mitigated.
- Efforts should be made to develop benchmarks to evaluate Gen AI in the P/CVERLT context. Benchmarks are commonly used in other high-stakes sectors (e.g., healthcare and law) to test Gen AI systems against predefined standards. In P/CVERLT, such benchmarks could evaluate whether a Gen AI model can, for

example, distinguish between non-violent extremism and incitement to terrorism. With combined policy and programmatic expertise in preventing and countering VERLT and the use of the internet for terrorist purposes, and supported by its regional focus and network of field operations, **the OSCE is well positioned to lead these efforts**. It can help reduce bias, misclassification, or misuse of Gen AI tools, ensuring that such tools in P/CVERLT are accurate, fair, and human rights-compliant

- **More funding should be allocated by participating States and the private sector to independent research by civil society and academia into the current use of Gen AI in P/CVERLT**, including on current practices across the sector, their effectiveness as well as what safeguards are in place to ensure that Gen AI tools are being used responsibly, ethically and legally (in terms of national and international law). This data should support multi-stakeholder efforts to develop good practice or guidance material (see below) on the use of Gen AI in P/CVERLT, grounded in human rights and do-no-harm principles

RECOMMENDATIONS FOR OSCE PARTICIPATING STATES AND CIVIL SOCIETY ENGAGEMENT WITH THE PRIVATE SECTOR

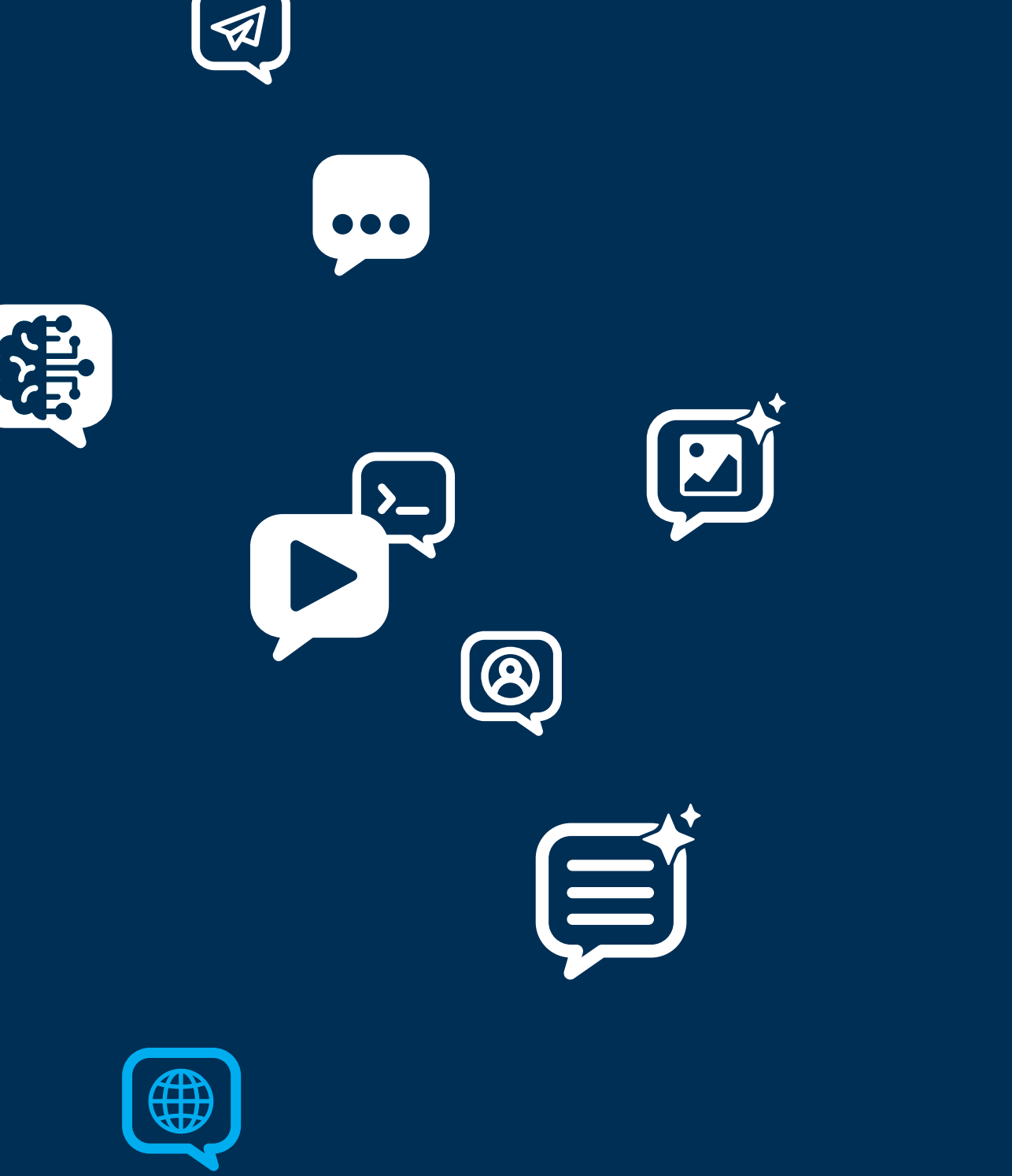
- **Governments and CSOs should ensure they possess the necessary digital literacy skills** — including a contemporary understanding of the Gen AI sector and the risks and challenges associated with the use of Gen AI tools — to engage with private sector actors selling Gen AI tools or solutions. P/CVERLT actors must be able to ask tool developers or providers the right questions – and understand the answers provided – before any procurement takes place, as well as be able to have an informed and ongoing dialogue if a tool is subsequently used.
- **Governments and CSOs should endeavour to engage with private sector actors in support of the above activities**. This could include engagement through existing voluntary, multi-stakeholder fora (e.g., GIFCT or the Christchurch Call Foundation), and programmatic partnerships between online platforms and CSOs to explore Gen AI's potential capabilities in P/CVERLT.
- **Governments should consider a variety of regulatory approaches alongside efforts to foster voluntary collaboration** to try to counteract the actions (or inaction) of some of the Gen AI companies and online platforms engaged in multi-stakeholder fora, which are helping to perpetuate the broader harms associated with Gen AI's development.[64]

64 McQue, K., "AI is overpowering efforts to catch child predators, experts warn", *The Guardian*, 18 July 2024. Available at: https://www.theguardian.com/technology/article/2024/jul/18/ai-generated-images-child-predators

P/CVERLT DONORS/FUNDERS

- **Donors and funders should integrate similar policies** and include requirements that good practices on the use of Gen AI are integrated in the projects they fund and support.
- **Donors and funders should ensure the above recommendations have been considered before asking P/CVERLT actors to integrate Gen AI into their activities.** This is a particular concern given the potential cost reduction benefits offered by Gen AI, and the temptation for funders/donors to promote Gen AI as a way for P/CVERLT actors to "deliver more with less," or replace existing services.
- **Donors should support efforts to establish or expand levels of digital literacy across the P/CVERLT community, alongside knowledge on human rights and fundamental freedoms, with a particular focus on Gen AI.**[65] Stakeholders at all levels should be able to critically assess the utility of any Gen AI tools/techniques that they procure or use, identify potential risks and seek to mitigate these risks by putting appropriate safeguards in place, ensuring that Gen AI tools are used in an ethical and legal manner or, if the risk is too high, are not used.

65 One example of this initiative is the joint GIFCT-Hedayah training programme to enhance the skills and knowledge of policymakers and CSOs. For more information, see the Hedayah website. Available at: https://hedayah.com/programs/stop_ve_online/



BIBLIOGRAPHY

Accenture, "What is generative AI", *Accenture*, n.d.
Available at: https://www.accenture.com/us-en/insights/generative-ai

Agarwal, A., "How AI Can Revive a Love of Learning", *New York Times*, 7 December 2024.
Available at: https://www.nytimes.com/2024/12/07/special-series/artificial-intelligence-schools-education.html

Ayoub, N.F., Balakrishnan, K., Ayoub, M.S., Barrett, T.F., David, A.P. and Gray, S.T., "Inherent Bias in Large Language Models: A Random Sampling Analysis", *Mayo Clinical Proceedings Digital Health*, June 2024.
Available at: https://www.mcpdigitalhealth.org/article/S2949-7612(24)00020-8/fulltext

Bajwa, A., "Nations building their own AI models add to Nvidia's growing chip demand", *Reuters*, 29 August 2024.
Available at: https://www.reuters.com/technology/nations-building-their-own-ai-models-add-nvidias-growing-chip-demand-2024-08-29/

Booth, R., "Losing our voice? Fears AI tone-shifting tech could flatten communication", *The Guardian*, 11 December 2024.
Available at: https://www.theguardian.com/society/2024/dec/11/ai-tone-shifting-tech-could-flatten-communication-apple-intelligence

Brenner, S., "Comparative lit class will be first in Humanities Division to use UCLA-developed AI system", *UCLA Newsroom*, 4 December 2024.
Available at: https://newsroom.ucla.edu/stories/comparative-literature-zrinka-stahuljak-artificial-intelligence?taid=6755debeb8836e000133c355&utm_campaign=trueAnthem_manual&utm_medium=trueAnthem&utm_source=twitter

Bressan, S., Ebbecke, S., and Rahlf, L., "How Do We Know What Works in Preventing Violent Extremism? Evidence and Trends in Evaluation from 14 Countries", *Global Public Policy Institute*, July 2024.
Available at: https://gppi.net/assets/BressanEbbeckeRahlf_How-Do-We-Know-What-Works-in-Preventing-Violent-Extremism_2024_final.pdf

Cloudfare, "What is a large language model (LLM)?", *Cloudfare*, n.d.
Available at: https://www.cloudflare.com/en-gb/learning/ai/what-is-large-language-model/

Crosling, G., Atherton, G. and Azizan, S.N., "The importance of TVET students' critical and flexible thinking skills for AI competence", *United Nations Educational, Scientific and Cultural Organization (UNESCO)*, October 2023.
Available at: https://unevoc.unesco.org/up/TVET_students_AI_competence.pdf

Fang, X., Shangkun, C., Mao, M., Zhang, H., Zhao, M. and Zhao, X., "Bias of AI-generated content: an examination of news produced by large language models", *Nature*, 4 March 2024.
Available at: https://www.nature.com/articles/s41598-024-55686-2

Faverio, M. and Sidoti, O., "Teens, Social Media and Technology", *Pew Research Center*, 12 December 2024.
Available at: https://www.pewresearch.org/internet/2024/12/12/teens-social-media-and-technology-2024/

Gerlich, M., "AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking", *Centre for Strategic Corporate Foresight and Sustainability*, Swiss Business School, 3 January 2025.
Available at: https://www.mdpi.com/2075-4698/15/1/6

Gorbacheva, A., "No language left behind: How to bridge the rapidly evolving AI language gap", *United Nations Development Programme (UNDP) Kazakhstan*, 4 October 2023.
Available at: https://innovation.eurasia.undp.org/ai-language-gap/

Gray Widder, D. and Hicks, M., "Watching the Generative AI Hype Bubble Deflate", *Harvard Kennedy School*, 20 November 2024.
Available at: https://ash.harvard.edu/resources/watching-the-generative-ai-hype-bubble-deflate/

Haider, J., Soderstrom, K.R., Ekstrom, B. and Rodl, M., "GPT-fabricated scientific papers on Google Scholar: Key features, spread, and implications for pre-empting evidence manipulation", *Harvard Kennedy School Misinformation Review*, 3 September 2024.
Available at: https://misinforeview.hks.harvard.edu/article/gpt-fabricated-scientific-papers-on-google-scholar-key-features-spread-and-implications-for-preempting-evidence-manipulation/

Hedayah, "Program Using Digital Platforms to Prevent and Counter Violent Extremist Propaganda", *Hedayah*, n.d.
Available at: https://hedayah.com/programs/stop_ve_online/

Hulick, K., "Google now adds watermarks to all its AI-generated content", *Science News Explores*, 11 December 2024.
Available at: https://www.snexplores.org/article/google-ai-watermarks

Humphrys, S., "Analysis: How jihadists experimented with AI in 2024", *BBC Monitoring*, 12 November 2024.
Available at: https://monitoring.bbc.co.uk/product/b0002qiw

Kannan. P., "Improving Equity and Access to Non-English Large Language Models", *Stanford University Human-Centered Artificial Intelligence*, 22 April 2024.
Available at: https://hai.stanford.edu/news/improving-equity-and-access-non-english-large-language-models

Kelsey-Sugg, A. and Carrick, D., "AI hallucinations caused artificial intelligence to falsely describe these people as criminals", *ABC News*, 3 November 2024.
Available at: https://www.abc.net.au/news/2024-11-04/ai-artificial-intelligence-hallucinations-defamation-chatgpt/104518612

Kenens, A. and Ivanovic, J., "Using generative AI for crisis foresight", *United Nations Development Programme (UNDP) Europe and Central Asia*, 20 November 2024.
Available at: https://www.undp.org/eurasia/blog/using-generative-ai-crisis-foresight

Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepano, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J. and Tseng, V., "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models", *PLOS Digital Health*, 9 February 2023.
Available at: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000198#:~:text=In%20this%20study%2C%20we%20evaluate,established%2C%20showing

Macdonald, S., Mattheis, A. and Wells. D., "Using Artificial Intelligence and Machine Learning to Identify Terrorist Content Online", *Tech Against Terrorism Europe*, 17 January 2024.
Available at: https://static1.squarespace.com/static/63e0c75f41ff767f07530a6f/t/65a5349648ce18428d686126/1705325720146/TATE+-+AI+REPORT+FINAL+%281%29.pdf

Martino, M., "Artificial intelligence is flooding the internet with fake images, video and audio. Can you tell real from fake?", *ABC News*, 13 September 2024.
Available at: https://www.abc.net.au/news/2024-09-14/artificial-intelligence-real-fake-quiz-abc-news-verify/104148236

Maxwell, T., "AI Chatbots Can Be Jailbroken to Answer Any Question Using Very Simple Loopholes", *Gizmodo*, 20 December 2024.
Available at: https://gizmodo.com/ai-chatbots-can-be-jailbroken-to-answer-any-question-using-very-simple-loopholes-2000541157

McKinsey & Company, "What are AI guardrails?", *McKinsey & Company*, 14 November 2024.
Available at: https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-are-ai-guardrails

McQue, K., "AI is overpowering efforts to catch child predators, experts warn", *The Guardian*, 18 July 2024.
Available at: https://www.theguardian.com/technology/article/2024/jul/18/ai-generated-images-child-predators

Microsoft, "Generative AI vs. other AI types", *Microsoft AI*, n.d.
Available at: https://www.microsoft.com/en-us/ai/ai-101/generative-ai-vs-other-types-of-ai#:~:text=,feedback%20loops%2C%20the%20system%20or

Milmo, D., "'Impossible' to create AI tools like ChatGPT without copyrighted material, OpenAI says", *The Guardian*, 8 January 2024.
Available at: https://www.theguardian.com/technology/2024/jan/08/ai-tools-chatgpt-copyrighted-material-openai?CMP=Share_AndroidApp_Other

Milosevic, A., "Testimonials: Involving victims/survivors of terrorism in P/CVE", *Radicalisation Awareness Network (RAN)*, 22 September 2023.
Available at: https://home-affairs.ec.europa.eu/document/download/b0c7b414-6837-4fa2-8599-7a6ad2ca056f_en?filename=ran_testimonials_involving_vot_in_pcve_092023_en.pdf

Mishcon de Reya, "Case tracker: Generative AI – Intellectual property cases and policy tracker", 26 June 2025.
Available at: https://www.mishcon.com/generative-ai-intellectual-property-cases-and-policy-tracker

Moorosi, N., "Better data sets won't solve the problem – we need AI for Africa to be developed in Africa", *Nature*, 11 December 2024.
Available at: https://www.nature.com/articles/d41586-024-03988-w

Mucci, T., "What is AI-generated content?", *International Business Machines Corporation (IBM)*, 27 November 2024.
Available at: https://www.ibm.com/think/insights/ai-generated-content

Organisation for Economic Co-operation and Development (OECD), "Generative AI", *OECD*, n.d.
Available at: https://www.oecd.org/en/topics/generative-ai.html

Organization for Security and Co-operation in Europe (OSCE), "Concept note 'Facing Division: Preventing and Countering Violent Extremist and Terrorist Content Online: Human rights-, age- and gender-sensitive approaches'", *OSCE*, 2024

Organization for Security and Co-operation in Europe (OSCE), "INFORMED: Information and Media Literacy in Preventing Violent Extremism – Human rights-based and gender-sensitive approaches to addressing the digital information disorder", *OSCE*, n.d.
Available at: https://www.osce.org/project/INFORMED

Organization for Security and Co-operation in Europe (OSCE), "OSCE enhances media and information literacy skills to effectively prevent and counter violent extremism", *OSCE*, 18 December 2024.
Available at: https://www.osce.org/secretariat/583642

Organization for Security and Co-operation in Europe (OSCE), "Summary Document of Expert-Level Event: Artificial Intelligence in the Context of Preventing and Countering Violent Extremism and Terrorism: Challenges, Risks and Opportunities", *OSCE*, n.d.
Available at: https://www.osce.org/files/f/documents/4/f/575877.pdf

Organization for Security and Co-operation in Europe (OSCE), "Strengthening Media and Information Literacy in the Context of Preventing Violent Extremism and Radicalization that Lead to Terrorism: A Focus on South-Eastern Europe", *OSCE*, September 2024.
Available at: https://www.osce.org/files/f/documents/4/4/575970.pdf

Peredaryenko, M. and Heng, A., "Beyond bullets: Leveraging AI and VR in P/CVE efforts", *Awani International*, 13 December 2024.
Available at: https://international.astroawani.com/malaysia-news/columnist-beyond-bullets-leveraging-ai-and-vr-pcve-efforts-500455

Perrigo, B., "Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic", *Time*, 18 January 2023.
Available at: https://time.com/6247678/openai-chatgpt-kenya-workers/

Pierson, B., "Mother sues AI chatbot company Character.AI, Google over son's suicide", *Reuters*, 24 October 2024.
Available at: https://www.reuters.com/legal/mother-sues-ai-chatbot-company-characterai-google-sued-over-sons-suicide-2024-10-23/

Popova Zhuhadar, L., "A Comparative View of AI, Machine Learning, Deep Learning, and Generative AI", *Wikimedia Commons*, 30 July 2023.
Available at: https://commons.wikimedia.org/wiki/File:Unraveling_AI_Complexity_-_A_Comparative_View_of_AI,_Machine_Learning,_Deep_Learning,_and_Generative_AI.jpg

Radicalisation Awareness Network (RAN), "Current challenges and solutions related to working with youth on P/CVE", *RAN Practitioners*, 1 December 2022.
Available at: https://home-affairs.ec.europa.eu/document/download/4cee01e1-94d3-46c7-8988-e57b0c2c0e85_en?filename=current_challenges_solutions_related_to_working_youth_on_pcve_22092022_en.pdf

Radicalisation Awareness Network (RAN), "Digital frontrunners: Key challenges and recommendations for online P/CVE work", *RAN Practitioners*, 16-17 June 2022.
Available at: https://home-affairs.ec.europa.eu/system/files/2022-08/ran_cn_digital_frontrunners_riga_16-17062022_en.pdf

Rahlf, L., "Preventing and Countering Violent Extremism in Europe: Expert Views on Contemporary Challenges", *VORTEX*, 3 July 2024.
Available at: https://vortex.uni.mau.se/2024/07/preventing-and-countering-violent-extremism-in-europe-expert-views-on-contemporary-challenges/

Robert, J., "Chatbot: definition, uses and impact on companies", *DataScientest*, 4 November 2024.
Available at: https://datascientest.com/chatbot-tout-savoir

Robison, K., "OpenAI is charging $200 a month for an exclusive version of its o1 'reasoning' model", *The Verge*, 5 December 2024.
Available at: https://www.theverge.com/2024/12/5/24314147/openai-reasoning-model-o1-strawberry-chatgpt-pro-new-tier

Sample, I., "Most AI chatbots easily tricked into giving dangerous responses, study finds", *The Guardian*, 21 May 2025.
Available at: https://www.theguardian.com/technology/2025/may/21/most-ai-chatbots-easily-tricked-into-giving-dangerous-responses-study-finds

Siegel, D., "'RedPilled AI': A New Weapon for Online Radicalisation on 4chan", *G-NET Insights*, 7 June 2023.
Available at: https://gnet-research.org/2023/06/07/redpilled-ai-a-new-weapon-for-online-radicalisation-on-4chan/

Stanley, J., "AI Generated Police Reports Raise Concerns Around Transparency, Bias", *American Civil Liberties Union*, 10 December 2024.
Available at: https://www.aclu.org/news/privacy-technology/ai-generated-police-reports-raise-concerns-around-transparency-bias

Tao. Y., Viberg, O., Baker, R.S. and Kizilcec, R.F., "Cultural bias and cultural alignment of large language models", *PNAS Nexus* 3(9), 17 September 2024.
Available at: https://academic.oup.com/pnasnexus/article/3/9/pgae346/7756548

Taylor, C., "How much is AI hurting the planet? Big tech won't tell us", *Mashable*, 3 September 2024.
Available at: https://mashable.com/article/ai-environment-energy

Ungoed-Thomas, J. and Abdulahi, Y., "Warnings AI tools used by government on UK public are 'racist and biased'", *The Guardian*, 25 August 2024.
Available at: https://www.theguardian.com/technology/article/2024/aug/25/register-aims-to-quash-fears-over-racist-and-biased-ai-tools-used-on-uk-public#:~:text=3%20months%20old-,Warnings%20AI%20tools%20used%20by%20government,public%20are%20'racist%20and%20biased'&text=Artificial%20intelligence%20and%20algorithmic%20tools,%E2%80%9Centrenched%E2%80%9D%20racism%20and%20bias

United Nations Educational, Scientific and Cultural Organization (UNESCO), "Challenging systematic prejudices: an investigation into bias against women and girls in large language models", *UNESCO*, 7 March 2024.
Available at: https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes

United Nations Educational, Scientific and Cultural Organization (UNESCO), "Recommendation on the Ethics of Artificial Intelligence", *UNESCO*, 23 November 2021.
Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137

Walsh, D., "MIT Study: An AI chatbot can reduce belief in conspiracy theories", *Massachusetts Institute of Technology (MIT) Management Sloan School*, 30 September 2024.
Available at: https://mitsloan.mit.edu/ideas-made-to-matter/mit-study-ai-chatbot-can-reduce-belief-conspiracy-theories#:~:text=The%20artificial%20intelligence%2Dpowered%20%E2%80%9CDebunkBot,according%20to%20a%20new%20study

Wells, D., "The Next Paradigm Shattering Threat? Right-sizing the Potential Impacts of Generative AI on Terrorism", *Middle East Institute*, 18 March 2024.
Available at: https://www.mei.edu/sites/default/files/2024-03/Wells%20-%20The%20Next%20Paradigm-Shattering%20Threat%20Right-Sizing%20the%20Potential%20Impacts%20of%20Generative%20AI%20on%20Terrorism.pdf

Wichmann, F., "Harnessing the Power of Victims, Survivors, and Former Extremists in Prevention Campaigns", *Journal EXIT-Deutschland*, 4 September 2023.
Available at: https://journal-exit.de/inclusive-voices-unite/

Williams, R. and O'Donnell, J., "We finally have a definition for open-source AI", *MIT Technology Review*, 22 August 2024.
Available at: https://www.technologyreview.com/2024/08/22/1097224/we-finally-have-a-definition-for-open-source-ai/

Wong, M., "The AI Revolution Is Crushing Thousands of Languages", *The Atlantic*, 12 April 2024.
Available at: https://www.theatlantic.com/technology/archive/2024/04/generative-ai-low-resource-languages/678042/

Organization for Security and
Co-operation in Europe