SUMMARY DOCUMENT OF EXPERT-LEVEL EVENT

Artificial Intelligence in the Context of Preventing and Countering Violent Extremism and Terrorism: Challenges, Risks and Opportunities

BACKGROUND

On 14 March 2024, the OSCE Transnational Threats Department/Action against Terrorism Unit (TNTD/ATU) in co-operation with the Office of the OSCE Representative on the Freedom of the Media (RFoM) and the International Centre for Counterterrorism (ICCT) held a technical expert meeting that aimed to:

- Discuss pressing issues of concern around challenges and opportunities associated with the use of artificial intelligence (AI) in the context of preventing and countering violent extremism and terrorism, with the aim of providing human rights-based and gender-sensitive policy guidance to OSCE participating States;
- Explore relevant topics for an upcoming OSCE webinar series related to P/CVERLT, AI and emerging technologies, that will form the basis for developing initial policy recommendations on AI and AI literacy in P/CVERLT;
- Support the development of an e-learning course “Using open-source intelligence tools to enhance Media and Information Literacy (MIL) skills among non-law enforcement actors, in the context of P/CVERLT”.

The following are a summary of the OSCE’s key findings from the event’s three sessions:

1) ARTIFICIAL INTELLIGENCE AND PREVENTION OF VIOLENT EXTREMISM AND RADICALIZATION THAT LEAD TO TERRORISM – CHALLENGES AND OPPORTUNITIES

MIL can help create resilience to violent extremist and terrorist narratives, both on the individual and societal level, and build awareness on the use of AI in generating content and propaganda for violent extremist and terrorist purposes, and how this is often gendered. There is a need for robust policies to tackle harmful content in the online space, in parallel with capacity-building on digital literacy amongst P/CVERLT stakeholders as well as the public.

Key findings:

- AI is improving capacities to create sophisticated, regionally- and locally-tailored propaganda.
- Enhancing social resilience, promoting digital literacy, and implementing technical AI countermeasures like watermarking* can help prevent the spread of false narratives and disinformation.
- MIL as an approach to P/CVERLT is increasingly recognized amongst the global counterterrorism and MIL communities.
- Generative AI and other new technological tools are leveraged by violent extremist and terrorist groups across the ideological spectrum to produce propaganda, including gender-based propaganda that is tailored to men and women, girls and boys. The exploitation of gendered narratives in general is an essential part of contemporary violent extremist and terrorist ideologies, often serving as a primary vector of radicalization to violence. The spread of this messaging must be considered to meaningfully address and counter the appeal of violent extremist narratives online, including when facilitated by new technologies.

*Watermarking refers to adding a pattern or logo embedded in AI-generated content that can help distinguish between AI- and human-generated content.

- AI is a tool that serves as an amplifier of existing issues and messages, and can be used to spread information at an accelerated pace that humans cannot. Responding to AI-generated extremist or terrorist content with takedowns alone is not realistic nor will it be efficient. Efforts to prevent VERLT messaging through the use of AI need to include critical thinking and enhanced digital skills.
- The ability of content creators to produce aesthetically appealing material has the ability to pull individuals, and especially youth, into increasingly violent content through ‘awful but lawful’ material. This may be by direct intent of the content creator, or through algorithmic amplification of content that increases user engagement even if it has a violent ideology behind it.
- A preventative approach to VERLT in the digital space needs to encompass social inclusion as experts observed that disenfranchised youth in particular may find content that is professionally made to look attractive with aesthetically pleasing videos and photos, but which is still but intolerant and hateful content attractive.
- Prevention should encompass people of all ages, but youth in particular need to be aware of the context and risks of online interactions. Everyone in society should be equipped with digital literacy skills to navigate online landscapes in a safe and positive way.
- Robust online harm policies both by states and internet intermediaries and human rights-based AI tools supporting human moderators, along with fostering collaboration through shared tools, and enhancing social resilience and digital literacy efforts are needed.
-

2) LAW ENFORCEMENT USE OF ARTIFICIAL INTELLIGENCE IN COUNTERING TERRORISM – ENSURING A HUMAN RIGHTS-BASED APPROACH

Law enforcement in many areas already make active use of AI tools in their work. This raises concerns about the negative impact of AI use for counterterrorism purposes on human rights, including privacy and fundamental freedoms, which need to be carefully considered and addressed by law enforcement and other stakeholders in order to minimize harm and promote trust and inclusion.

Key findings:

- Tools such as INTERPOL's toolkit for responsible use of AI in law enforcement underscores the need for human oversight, which also addresses data curation, community trust, and transparency.
- All AI systems are biased to some degree, and they reflect existing and historical societal biases. The use of technology in counterterrorism should therefore be carefully assessed in light of the historical context of counterterrorism and its specific biases, in particular in light of issues related to the opaqueness of algorithms and data analysis which can lead to adverse effects on human rights.
- Risk assessment on the use of AI in law enforcement need to be carried out regularly and approaches adjusted accordingly, in order to center law enforcement policies and practices on human rights and fundamental freedoms.
- The organizational culture on the use of technology needs to be considered, especially given the historical context of counterterrorism and its specific biases, which means addressing internal digital literacy levels amongst the users of the technology. This includes awareness-raising and capacity-building, but also evaluation and harm mitigation on an ongoing basis.

- There are various human rights and ethical implications of the use of AI in decision-making processes that need to be addressed, often identified by decline in transparency and accountability.
- Development and evaluation of AI use by law enforcement can be strengthened by the inclusion of diverse perspectives, including from civil society and human rights experts.

3) FREEDOM OF EXPRESSION, ARTIFICIAL INTELLIGENCE AND (PREVENTING) VIOLENT EXTREMISM

P/CVERLT cannot be sustainable if it comes at the expense of freedom of expression. Effectively addressing VERLT online requires an analysis of the adverse effects of violent extremist and terrorist groups' use of AI on the digital democratic space. It also means identifying how, and when, AI can be leveraged to foster healthy online environments without negatively impacting freedom of expression.

Key findings:

- The role of social media companies in detecting and acting upon harmful content is evident but is often limited to 'remove or remain', with limited use of less restrictive approaches such as 'quarantining' or 'downranking' options. If done in a transparent, non-biased and accountable manner, such compromise options can be useful instead of removing content that is not illegal in nature, considering the potential adverse impacts this type of harmful content can have on freedom of expression and radicalization to violence.
- Despite rapid developments of AI technology, current limitations of AI-powered content moderation – such as its inability to understand linguistic, cultural, or political contexts – make ongoing human review and process involvement necessary in addressing both illegal and so-called 'borderline' content.
- AI bias disproportionately impacts marginalized groups, demonstrating the urgent need for meaningful human rights due diligence, accountability and algorithmic transparency by online platforms.
- While social media platforms have a distinct responsibility to remove terrorist content online, there is also a need to ensure preservation of content that may have evidential value.

- There is a need for increased transparency regarding the criteria for designating dangerous organizations, including terrorist organizations, by social media platforms.
- Uneven resource allocation to content moderation of different languages and in different regions causes disproportionate and inadequate moderation in some languages which can cause an increasingly negative impact on some regions as well as on minority populations.
- Individual accountability in the form of efficient appeals mechanisms is essential for preserving access to rights for users whose content is removed.
- Regulatory frameworks are new and often limited, if not non-existent. There is an urgent need for human rights-based regulation of the use of AI in the online space in order to ensure sustainable P/CVERLT. Such regulation should address both content governance and broader platform governance issues, such as interlinkages with targeted advertisement, power concentration, and surveillance-based business models.
- Awareness of the right to freedom of expression in the digital space needs to be increased, and adequate content moderation approaches should come alongside strategies to tackle disinformation, such as MIL.

CONCLUSION

The key findings demonstrate that AI and other new technologies need to be addressed with a risk- and rights-based approach. In P/CVERLT and counterterrorism efforts this includes thorough analyses of the ways violent extremist and terrorist groups exploit technology for recruitment, propaganda and messaging. While the technology is novel, the way it is used by violent extremist and terrorist groups reflects the ideologies and mindset that have existed on- and offline for many years, which includes the exploitation of gendered narratives and propaganda. VERLT is often interlinked with disinformation, which raises specific challenges related to the speed, scale and scope that this can spread with the use of AI. Law enforcement must be especially cautious in their use of AI in counterterrorism efforts, in order not to deteriorate trust or infringe rights due to opaque decision-making processes as well as organizational capacity and knowledge of this technology. This cannot be separated from the historical biases of the counterterrorism efforts of the past decades. A rights-based approach has to consider the impact that technology and (lack of) regulation has on freedom of expression in the online space, in particular on already marginalized groups. Sustainable strategies can only be found when online platforms and service providers are engaged alongside other stakeholders such as civil society, media, international actors and academia, which emphasizes the need for adequate regulation. Algorithmic transparency, human rights impact assessments, and efficient appeal mechanisms are some of the harm mitigation measures which platforms can provide.

Measures that remove or downgrade content are not going to be sufficient in the face of the sheer volume of information that is uploaded on a daily basis. Regulation should also ensure accessibility and visibility of quality and public interest content. Users of all ages need to be supported in approaching digital content with critical thinking, analytical skills and the ability to understand how AI impacts their online experience.

The OSCE has identified four focus areas in its work to address the use of AI and emerging technologies in P/CVERLT and counterterrorism:

1. Through the Extrabudgetary projects ‘INFORMED’ and ‘E-VIDENCE’.
2. The development of an e-learning course on “Using open-source intelligence tools to support media and information literacy skills within the law enforcement actors, in the context of P/CVERLT”.
3. An OSCE expert-level webinar series on AI and emerging technologies and their impact on P/CVERLT.
4. The OSCE RFoM is addressing AI, freedom of expression and media freedom through its ‘Healthy Online Information Spaces – SAIFE Renewed’ project.