Artificial Intelligence: Threats, Opportunities, and Policy Frameworks for Countering VNSAs

by **Erin Saltman & Skip Gilmour**

April 2025

KONRAD ADENAUER STIFTUNG
UN LIAISON OFFICE NEW YORK
GIFCT
Global Internet Forum to Counter Terrorism

KONRAD ADENAUER STIFTUNG | UN LIAISON OFFICE NEW YORK

GIFCT Global Internet Forum to Counter Terrorism

Artificial Intelligence: Threats, Opportunities, and Policy Frameworks for Countering VNSAs

By Erin Saltman and Skip Gilmour*[1]

Dr. Erin Saltman is the Senior Director of Membership & Programs at GIFCT. She has worked in the technology, NGO, and academic sectors, building out international counterterrorism strategies and programs. Her background and expertise span a range of regional and socio-political contexts. Her research and publications have focused on the evolving nature of violent extremism online, youth radicalization, and the evaluation of counterspeech approaches.

Skip Gilmour is the Director of Trust & Safety Solutions at GIFCT. He has worked in the intelligence community and technology sector investigating and developing solutions to mitigate counterterrorism threats. His past work and expertise includes a focus on racially and ethnically motivated violent extremism, youth radicalization, and the instrumentalization of gaming and online communities by terrorist and violent extremist actors.

About GIFCT: Mission and Work

The Global Internet Forum to Counter Terrorism (GIFCT) is a unique tech-led nonprofit organization dedicated to preventing terrorists and violent extremists from exploiting digital platforms. Founded in 2017, GIFCT convenes more than 30 technology platforms and fosters collaboration and dialogue with governments, civil society, practitioners, and academia to advance collective counterterrorism efforts through key tools that include the Hash[2] Sharing Database and Incident Response Framework. Through its academic arm, the Global Network on Extremism and Technology (GNET), GIFCT connects global experts with the tech sector, providing critical insights into emerging threats and trends shaping digital safety.[3]

About KAS

The Konrad-Adenauer-Stiftung (KAS) is a German political foundation closely associated with the Christian Democratic Union (CDU). With more than 100 offices worldwide, KAS advocates for democracy and the rule of law, the vision of a unified Europe, a social market economy, and a rules-based multilateral order. Its New York Office provides a platform for dialogue and cooperation between representatives of the United Nations system, Member States, experts and partners from its global network, and KAS offices around the globe, offering new perspectives on issues relating to peace and security, sustainable development, and global governance.

1 The views and opinions expressed in this policy brief are those of the authors and do not reflect the official policies or positions of GIFCT or the Konrad Adenauer Foundation.

2 Hashes: A unique string of characters that corresponds to a specific piece of media. Sometimes called "digital fingerprints", hashes are used to identify content online.

3 For more information referring to GIFCT's work mitigating AI risk see appendix 3.

Introduction

The following policy brief was developed by the Global Internet Forum to Counter Terrorism (GIFCT) in partnership with the Konrad-Adenauer-Stiftung (KAS) to examine the intersections of artificial intelligence (AI) with non-state actor, terrorist and violent extremist exploitation online and how best to frame policy to ensure safety by design. This policy brief was informed by insights on AI developed by GIFCT's academic wing, the Global Network on Extremism and Technology (GNET), as well as discussions with industry representatives, global experts, and civil society. To that end, GIFCT and KAS co-hosted a virtual dialogue with 40 international researchers, practitioners, tech company representatives, and United Nations (UN) participants reviewing case studies of non-state actor, terrorist and violent extremist exploitation of generative artificial intelligence (GenAI). GIFCT also held bilateral conversations with tech companies deploying GenAI products for public use and safety efforts to inform this paper, and analyzed 13 multi-stakeholder AI governance frameworks.[4]

The emergence of technologies like GenAI has heightened concerns about their manipulation and utilization by violent non-state actors (VNSAs), and in particular, terrorist and violent extremist groups. Such groups have historically exploited new technologies to further their goals and ideas, and research has shown that adversarial actors have sought to utilize GenAI to create propaganda, further their operational endeavors, and sow confusion in the information environment. The proliferation of GenAI usage by average online users in the last two years has also driven a wave of voluntary policy and legislative frameworks around the world focused on AI in an attempt to ensure safety concerns can be met.

Following a brief overview of the contemporary threat landscape and evolving trends, this paper aims to examine specific concerns as well as potential solutions, concluding with recommendations for a range of stakeholders. This brief aims to:

- Map cases where experts have documented exploitation by VNSAs using GenAI and emerging technologies,[5]
- Discuss where experts have concerns for further exploitation of new technologies,
- Map how GenAI is currently and could be used further to counter adversarial behavior and fueling terrorist and violent extremist content (TVEC) online,
- Outline policy recommendations for tech companies, policymakers, and practitioners with considerations to existing frameworks.

4 For a full list of analyzed frameworks see appendix 1.

5 This paper will largely focus on terrorist and violent extremist actors as a subset of VNSAs, recognizing that VNSAs covers a wider range of potential groups and individuals.

How Is AI Being Used by Terrorist and Violent Extremist Groups?

In essence, AI is the ability of a computer or a machine to think or learn. This includes a model or interface being able to engage and develop visual perception, speech recognition, decision-making, problem-solving, and translation tasks, among other things. GenAI is non-deterministic, meaning it produces variable responses to users, generating content based on growing and changing training data, enabling it to re-combine data in unique ways.

AI is increasingly deployed for everyday use by individuals around the world, via chatbots, search engines, and graphic design support. However, the advent of ChatGPT in 2022, and subsequent releases of competing user-facing GenAI models, have made large language models (LLMs) accessible to everyday internet users, enhancing their abilities to access and generate content. At the same time, image generation models enable users to create image content and multi-media through simple text description. Like other emerging technologies, these tools have enormous potential but can lead to harm when deployed by malicious actors, including VNSAs such as terrorist and violent extremist groups.

While the technology is still relatively new, researchers and experts have already highlighted a few emerging trends. Some of these trends could be considered higher prevalence but potentially lower offline harm risk, while others may be lower prevalence online but pose higher risks for offline harm.

Exploitation of GenAI and AI Enhanced Threats by VNSAs

A number of experts and practitioners have noted that VNSAs are now able to more easily scale the quality and quantity of propaganda and TVEC through GenAI products (Engler, 2023; Borgonovo et. al., 2024; Dean, 2025). To identify risks and gaps, practitioners are attempting to "jailbreak"[6] GenAI models, mimicking how bad actors might try to circumvent safety efforts. Increasingly, tech companies also deploy "red teaming" efforts, whereby a company proactively engages experts to show them where loopholes in their safety measures might exist. Among the concerns highlighted by researchers and practitioners are:

Content Moderation Evasion: GenAI tools are being used to augment and manipulate TVEC to evade social media platform detection and content moderation efforts.

- For example, overlaying livestream attacker footage, such as the Christchurch attacks in 2019, with "minions" cartoon character images over the victims to evade detection and gamify the attack (AI Virtual Dialogue, 2025).

6 Jailbreaking is a technique, or series of techniques, aimed at breaking the guardrails or mitigations in place within a GenAI model. Harmful content or results can come from finding ways to circumvent guardrails with the aim of causing the system to violate its own operation policies, with the aim of executing malicious instructions.

Propaganda and Disinformation: Contextually tailored and personalized propaganda has become easier to create. This also includes "deepfakes" of real or fictional online personalities, audio re-creations, and voice-overs to create more compelling content. VNSAs have also been documented modifying existing games online or creating their own online game-play spaces with concern that game creation engines could make increasingly compelling propaganda (Lamphere-Englund, 2025).

- For example, more refined GenAI images are now seen across official ISIS-affiliated publications (Borgonovo et. al., 2024).
- For example, genre-specific anti-Semitic and racist pop music with lyrics to weaponize and influence listeners has been seen in the UK linked to recent VNSA riots in August 2024 (Lopes, 2024).

Recruitment and Radicalization: Experts have raised concerns that GenAI can strengthen the abilities for VNSAs to personalize recruitment efforts, particularly given the importance placed on tailored and targeted strategies by groups like ISIS. Recruitment chatbots could be deployed in social media and gaming environments online in increasingly effective ways, mimicking everyday conversations, nudging curious users towards more extreme content.

- For example, AI researchers have used jailbreak tactics combining chatbot and search engine functions to share information leading to an al-Qaeda website and share a link to the repository of al-Qaeda propaganda productions (AI Virtual Dialogue, 2025; Siegal, 2025).
- For example, deepfakes of the Bali bombers have emerged in Indonesia, re-animating the attackers to appear to be telling audiences to carry out attacks and inciting violence (AI Virtual Dialogue, 2025).

Attack Planning: AI tools can facilitate innovative and efficient ways of planning and operationalizing attacks by making critical information widely available. Additionally, concerns have been flagged about the usage of chatbots and cyber criminal tactics to raise funds through cryptocurrency. Experts also flagged concerns about the circulation of information regarding bomb making, 3D printed firearms, and future potentials of chemical and biological weapon design (See also: GIFCT UNGA Side Event, 2024).

- For example, in the USA, there is a "3D2A" movement online combining 3D printing with Second Amendment (2A) rights, guiding VNSAs on how to use GenAI to facilitate partial or fully 3D printed assault weapons (3D2A GNET Insight, 2025).
- For example, in East Africa, groups like Al Shabaab have been documented using drones for reconnaissance and to record footage for propaganda (Aguilera, 2023; Figueiredo, 2024).

Attack Operationalization: While the full deployment of AI-facilitated attack capabilities by VNSAs has not yet been seen, it has been deployed in other military and armed conflict areas and could spill into VNSAs exploitation in the near future (AI Virtual Dialogue, 2025).

- For example, in Ukraine, we have seen the implementation of "smart drones" using AI assistance. There is fear of VNSAs scaled usage of AI enhanced drones for attacks.
- For example, following the January 2025 explosion of a Tesla Cybertruck outside of the Trump Hotel in Las Vegas, many feared the exploitation of autonomous driving vehicles for terrorist attacks, particularly in light of the increased usage by ISIS of van and car rammings into crowded areas.

Content produced by known terrorist networks would violate the policies of GIFCT members, regardless of whether the content was violent or not. However, it can be increasingly difficult to proactively detect potentially harmful networks and their content without awareness of how jailbreaking is taking place or without knowledge of nuanced global trends on how VNSA content manifests, particularly when not all violent extremist groups are on government designation lists. As the above examples of VNSA exploitation show, there are large variations in higher prevalence but lower risk GenAI content production, such as the ability to make better propaganda, in comparison to lower prevalence but higher risk exploitation of online tools to create and deploy weapons or cause real-world harm.

Tech Sector Mitigation Efforts for AI Enhanced Threat of VNSAs

The tech sector has proactively driven efforts to develop safeguards and tools to mitigate challenges posed by VNSA exploitation of AI. These challenges include content-hosting platforms seeking to detect AI generated harms at scale, and AI providers moderating an AI system's potential output. For the former, many established VNSA mitigations remain applicable to AI generated harms. In the latter circumstance, AI content generation's emergent nature necessitates novel strategies to categorize threat areas and govern how AI may respond to a user's prompt. In either case, AI's impact on the tech sector's efforts to mitigate VNSAs cannot be understated. GIFCT has collaborated with members on these efforts, and continues to facilitate solutions to stop the spread of harmful content, agnostic of how the content was generated.

How Can AI Help Platforms Address the Threat from VNSAs?

VNSAs are constantly evolving and adapting their tactics, often in direct response to platform efforts to block or mitigate their previous strategies. This ongoing cycle of adaptation underscores the need for platforms to deploy advanced AI tools capable of outpacing and countering the increasingly sophisticated methods used by these actors. Fortunately, the tech sector has been pioneering techniques to improve moderation capabilities using ML for years.

The following are examples of AI assisted tools used by platforms for scaled moderation of policy-violating content, which includes, though is not limited to:

- **Logo Recognition**: An application of computer vision used to identify and discover content, which includes logos related to bad actor networks.
  - E.g., Logo recognition could identify VNSA organization logos or commonly repeated motifs in an image.
- **Text Classification**: An application of natural language processing for categorizing text into pre-defined labels based on its content, linguistic features, and contextual signals.
  - E.g. Text classification could identify coded language or recruitment efforts used by a VNSA.
- **Alternate Account Detection**: Tooling that can algorithmically assess the likelihood that two accounts are operated by the same user.
  - E.g., Alternate Account Detection could be used to detect recidivist VNSAs returning to a platform after being banned.

Using these and other methods, platforms are able to quickly moderate overt policy violations, identify novel ban evasion techniques, and efficiently route complex abuse to human reviewers. Efforts to train AI for the purposes of VNSA enforcement should accompany persistent, contextual data collection overseen by subject matter experts to achieve ideal results.

How do AI Providers Approach the Problem of VNSA Exploitation?

Interest in AI systems that are capable of complex content generation has grown rapidly amongst the general public and VNSAs alike. To minimize risk, AI providers proactively implement safety features targeted at critical harms like VNSA exploitation. These efforts aim to stop GenAI systems from outputting dangerous content. According to discussions with GIFCT partners, many AI providers develop strategies or taxonomies to categorize harms. These strategies organize efforts to mitigate vulnerabilities. This is often done by red teaming, which involves human specialists simulating adversarial (potentially VNSA) efforts to exploit AI.

The findings from these efforts are used to inform AI safety features, including but not limited to:

- **Prompt Filtering**: A pre-process safety feature that evaluates prompts before they are given to a GenAI model.
  - E.g., Prompt filtering could identify violent language in a VNSA's prompt, and alter it before sending it to the GenAI model to neutralize risk.

- **Contextual Guardrails**: A model-level safety feature that enables the AI to respond to prompts based on the context of previous prompts.
  - E.g., Contextual guardrails could identify patterns in a VNSA's prompt that indicate an effort to jailbreak the AI. A VNSA may seek to do this to make the AI produce a recipe for explosives.

These efforts to implement GenAI safety features underline the importance of maintaining subject matter expertise on VNSAs and other harm areas to effectively detect vulnerabilities.

Multi-Stakeholder Recommendations: Mitigating VNSA Exploitation of AI Together

Key Recommendation: Strengthen Collaboration

Multi-stakeholder collaboration is key to successfully mitigating risks posed by VNSAs seeking to exploit AI. GIFCT has identified opportunities for further multi-stakeholder engagement and research, including:

- An exploration of how international and non-standardized VNSA definitions impact AI safety strategies and overall safety outcomes.
- Research measuring the impact of VNSA exploitation of AI compared to similar VNSA activity executed with non-AI tools.

Tech Sector Recommendations: VNSA Industry Standards

Tech sector partners should share AI safety strategies and successful techniques for countering VNSA exploitation of AI. This may enable smaller or less experienced AI developers to more effectively mitigate VNSA exploitation. GIFCT has identified additional opportunities for sharing, including:

- Common VNSA strategies/taxonomies: Tech sector partners should consider collaborating to create a common understanding of how different VNSAs interact with AI systems across various harm types. This collaboration may enable smaller AI developers to implement successful strategies with less investment cost.

Public Sector Recommendations: Continued Partnership with SMEs

The public sector should continue to partner with subject matter experts, civil society, academics, and the tech sector. GIFCT has identified opportunities for further public sector engagement, including:

- Working with subject matter experts to craft guidance for regulatory compliance for VNSA exploitation of AI, including how AI providers should assess causality when reporting violence related to AI outputs.
- Partnering with subject matter experts to maintain awareness of VNSA trends, which are key to identifying and mitigating vulnerabilities in GenAI systems.

Appendix 1

GIFCT assessed the following AI frameworks, commitments, and regulations to inform our analysis. We prioritized reviewing a variety of global standards with varying focuses. Many frameworks were aimed at providing guidance to policymakers. Some frameworks suggested guidelines for AI providers, and a small number included obligations for AI actors to follow.

AI Frameworks, Commitments, and Regulations			
Name	Entity	Type	Focus
Roles and Responsibilities Framework for Artificial Intelligence in Critical Infrastructure	DHS (USA)	Voluntary	AI and Critical Infrastructure
Interim Measures for the Administration of Generative Artificial Intelligence Services	CAC (CN)	Regulatory	AI Product Regulation
AI Act	EU	Regulatory	AI Product Regulation
Frontier AI Safety Commitments, AI Seoul Summit 2024	UK, KR	Voluntary	General AI Product Safety
The Voluntary AI Safety Standard	AU	Voluntary	General AI Product Safety
Hiroshima Process International Code of Conduct	G7	Voluntary	General AI Product Safety
Artificial Intelligence Risk Management Framework	NIST (USA)	Voluntary	General AI Product Safety
Framework Convention on AI and Human Rights, Democracy, and the Rule of Law	Council of Europe	Voluntary	Governance Approach
Framework for the Classification of AI Systems	OECD	Voluntary	Governance Approach
Governing AI for Humanity	HLAB (UN)	Voluntary	Governance Approach
Readiness Assessment Methodology	UNESCO (UN)	Voluntary	Governance Approach
Continental Artificial Intelligence Strategy	African Union	Voluntary	Governance Approach
Resolution A/78/L.49	UNGA (UN)	Voluntary	Governance Approach

Appendix 2

GIFCT and GNET References and Recommended Reading

The Global Network on Extremism and Technology (GNET) is the academic wing of GIFCT. GNET Insights provide short, action-oriented briefings on topics related to extremism and technology from experts around the world. The below insights and GIFCT resources are related specifically to AI threats and opportunities and are all publicly accessible.

Agnolon, A. 28 Jan 2025. AI Tools and the Alt-Right: A Double-Edged Sword for P/CVE. GNET Insights. https://gnet-research.org/2025/01/28/ai-tools-and-the-alt-right-a-double-edged-sword-for-p-cve/

Aguilera, A. 5 July 2023. Drone Use by Violent Extremist Organisations in Africa: The Case of Al-Shabaab. GNET Insights. https://gnet-research.org/2023/07/05/drone-use-by-violent-extremist-organisations-in-africa-a-case-study-of-al-shabaab/

Anadi. 11 Sept 2024. Deep Fakes, Deeper Impacts: AI's Role in the 2024 Indian General Election and Beyond. GNET Insights. https://gnet-research.org/2024/09/11/deep-fakes-deeper-impacts-ais-role-in-the-2024-indian-general-election-and-beyond/

Borgonovo, F., Bolpagni, A., Rizieri Lucini, S. 9 May 2024. AI-Powered Jihadist News Broadcasts: A New Trend In Pro-IS Propaganda Production? GNET Insights. https://gnet-research.org/2024/05/09/ai-powered-jihadist-news-broadcasts-a-new-trend-in-pro-is-propaganda-production/

Bowes, J. 20 Mar 2024. 'It's Over! White People are Finished': Accelerationist Memes using Generative AI on 4chan's '/pol'. GNET Insights. https://gnet-research.org/2024/03/20/its-over-white-people-are-finished-accelerationist-memes-using-generative-ai-on-4chans-pol/

Dean. L. 13 Jan 2025. AI or Aryan Ideals? A Thematic Content Analysis of White Supremacist Engagement with Generative AI. GNET Insights. https://gnet-research.org/2025/01/13/ai-or-aryan-ideals-a-thematic-content-analysis-of-white-supremacist-engagement-with-generative-ai/

Dean, L. 25 Feb 2025. AI or Aryan Ideals? Part Two: A Thematic Content Analysis of White Supremacist Engagement with Generative AI: Discourse. GNET Insights. https://gnet-research.org/2025/02/25/ai-or-aryan-ideals-part-two-a-thematic-content-analysis-of-white-supremacist-engagement-with-generative-ai-discourse/

Deedman, J. 30 Aug 2023. For the AI Generation, We Need Education as Much as Regulation. GNET Insights. https://gnet-research.org/2023/08/30/for-the-ai-generation-we-need-education-as-much-as-regulation/.

Ename Minko, A. 27 Sept 2024. AI Against Terror: Harnessing Technology to Combat Terrorism in the Horn of Africa. GNET Insights. https://gnet-research.org/2024/09/27/ai-against-terror-harnessing-technology-to-combat-terrorism-in-the-horn-of-africa/

Ename Minko, A. 23 Jan 2025. Leveraging AI-Driven Tools: Enhancing Moderation and Crisis Management on Smaller Digital Platforms in Africa. GNET Insights. https://gnet-research.org/2025/01/23/leveraging-ai-driven-tools-for-capacity-building-in-crisis-response-enhancing-moderation-and-crisis-management-on-smaller-digital-platforms-in-africa/

Engler, M. 20 Sept 2023. Considerations of the Impacts of Generative AI on Online Terrorism and Extremism: GIFCT Red Teaming Working Group. GIFCT Working Groups. https://gifct.org/wp-content/uploads/2023/09/GIFCT-23WG-0823-GenerativeAI-1.1.pdf

Figueireda, B. A. January 2024. The Use of Uncrewed Aerial Systems by Non-State Armed Groups: Exploring Trends in Africa. UNIDIR. https://unidir.org/wp-content/uploads/2024/01/UNIDIR_Use_of_Uncrewed_Aerial_Systems_by_Non_State_Armed_Groups_Africa.pdf

Klempner, U. and Koblentz-Stenzler, L. 11 Jul 2024. Methodologies in Manipulation: The Far-Right's Antisemitic Discourse Online Amid the Israel-Hamas War, GNET Insights. https://gnet-research.org/2024/07/11/methodologies-in-manipulation-the-far-rights-antisemitic-discourse-online-amid-the-israel-hamas-war/

Koblentz-Stenzler, L. and Klempner, U. 25 Jan 2024. Navigating Far-Right Extremism in the Era of Artificial Intelligence. GNET Insights. https://gnet-research.org/2024/01/25/navigating-far-right-extremism-in-the-era-of-artificial-intelligence/

Lakomy, M. 15 Dec 2023. Artificial Intelligence as a Terrorism Enabler? Understanding the Potential Impact of Chatbots and Image Generators on Online Terrorist Activities. GNET Insights. https://gnet-research.org/2023/12/15/artificial-intelligence-as-a-terrorism-enabler-understanding-the-potential-impact-of-chatbots-and-image-generators-on-online-terrorist-activities/

Lamphere-Englund, G. February 2025. 2024 Resource List: Violent Extremism, Radicalization, and Gaming. GIFCT Year 4 Working Group: Gaming Community of Practice. https://gifct.org/wp-content/uploads/2025/02/GIFCT-25WG-0225-EG-Resources-1.1.pdf

Lopes, H. 11 Dec 2024. Melodies of Malice: Understanding How AI Fuels the Creation and Spread of Extremist Music. GNET Insights. https://gnet-research.org/2024/12/11/melodies-of-malice-understanding-how-ai-fuels-the-creation-and-spread-of-extremist-music/

Mishra, A. and Karumbaya, V. 15 Feb 2024. A Deadly Trifecta: Disinformation Networks, AI Memetic

Warfare, and Deepfakes. GNET Insights. https://gnet-research.org/2024/02/15/a-deadly-trifecta-disinformation-networks-ai-memetic-warfare-and-deepfakes/

Nelu, C. 24 Feb 2025. Harnessing AI for Online P/CVE Efforts: Tools, Challenges, and Ethical Considerations. GNET Insights. https://gnet-research.org/2025/02/24/harnessing-ai-for-online-p-cve-efforts-tools-challenges-and-ethical-considerations/

Okechukwu Effoduh, J. 8 Mar 2024. The Role and Potential of Artificial Intelligence in Extremist Fuelled Election Misinformation in Africa. GNET Insights. https://gnet-research.org/2024/03/08/the-role-and-potential-of-artificial-intelligence-in-extremist-fuelled-election-misinformation-in-africa/

Risius, M., Namvar, M., Akhlaghpour, S., Xie, H. 24 Jun 2024. "Substitution": Extremists' New Form of Implicit Hate Speech to Avoid Detection. GNET Insights. https://gnet-research.org/2024/06/24/substitution-extremists-new-form-of-implicit-hate-speech-to-avoid-detection/

Saltman, E. and Johnson, S (eds.). 2023. Playbook on Positive Intervention Strategies Online: GIFCT Blue Team Working Group. GIFCT Working Groups. https://gifct.org/wp-content/uploads/2023/11/GIFCT-23WG-1023-Playbook-1.1.pdf

Shah, M. 4 Jul 2024. The Digital Weaponry of Radicalisation: AI and the Recruitment Nexus. GNET Insights. https://gnet-research.org/2024/07/04/the-digital-weaponry-of-radicalisation-ai-and-the-recruitment-nexus/

Siegal, D. 19 February 2024. AI Jihad: Deciphering Hamas, Al-Qaeda and Islamic State's Generative AI Digital Arsenal. GNET Insights. https://gnet-research.org/2024/02/19/ai-jihad-deciphering-hamas-al-qaeda-and-islamic-states-generative-ai-digital-arsenal/

Testa, P. 18 Mar 2024. CLASSIFIED 1948/2024: What Israeli AI Implementation Teaches Us About the Warfare of Tomorrow. GNET Insights. https://gnet-research.org/2024/03/18/classified-1948-2024-what-israeli-ai-implementation-teaches-us-about-the-warfare-of-tomorrow/

Thorley, T. and Saltman, E. May 2023. GIFCT Tech Trials: Combining Behavioural Signals to Surface Terrorist and Violent Extremist Content Online. Studies in Conflict & Terrorism. https://www.tandfonline.com/doi/full/10.1080/1057610X.2023.2222901

Zuroski, N. 6 Dec 2023. Weapon or Tool?: How the Tech Community Can Shape Robust Standards and Norms for AI, Gender, and Peacebuilding. GNET Insights. https://gnet-research.org/2023/12/06/weapon-or-tool-how-the-tech-community-can-shape-robust-standards-and-norms-for-ai-gender-and-peacebuilding/

8 Oct 2024. GIFCT Side Events at the 79th UNGA: Emerging Technologies, Counterterrorism Sanctions, and Youth PVE Initiatives. GIFCT News.
https://gifct.org/2024/10/08/news-gifct-side-events-at-the-79th-unga-emerging-technologies-counterterrorism-sanctions-and-youth-pve-initiatives/

6 February 2025. 3D2A: The Second Amendment, 3D Printed Guns and Memed Accelerationism. GNET Insights. https://gnet-research.org/2025/02/06/3d2a-the-second-amendment-3d-printed-guns-and-memed-accelerationism/

Appendix 3

How Has GIFCT Responded to GenAI and AI Enhanced Threat by VNSAs?

GIFCT's longstanding mission is to prevent terrorists and violent extremists from exploiting digital platforms, whether created by traditional means or via AI tools. The Hash-Sharing Database (HSDB), for example, serves as a hub for sharing hashed TVEC, and the GIFCT Incident Response Framework enables a rapid response to offline threats with an online nexus. GIFCT member companies have been exposed to augmented and synthetic TVEC designed to bypass moderation systems even before GenAI was widely used. That being said, there is widespread recognition that AI and GenAI in particular have changed the threat and opportunity landscape. As such, GIFCT continues to dedicate thematic workstreams to this topic, including:

- In 2023, GIFCT's "Red Team" and "Blue Team" Working Groups explored ways in which TVE groups had begun to utilize AI and what avenues existed for prevention and mitigation strategies using AI tools.[7]
- In 2024, GIFCT brought together industry leaders, policymakers, and practitioners in New York for an event on the margins of the UN General Assembly meetings to explore the impacts and implications of AI, unmanned aerial systems (UAS), and 3D printing on international counterterrorism cooperation.
- In 2025, GIFCT worked with KAS to convene a virtual dialogue with 40 international researchers, practitioners, tech company representatives, and UN participants reviewing case studies of non-state actor, terrorist and violent extremist exploitation of GenAI, feeding into this paper.
- In 2025, GIFCT launched a dedicated AI: Threats and Opportunities multistakeholder and cross-sector Working Groups to produce analysis and best practices.
- GIFCT's academic wing, the Global Network on Extremism and Technology (GNET), has had a dedicated focus on producing insights on the exploitation of AI by TVE actors, as well as highlighting where new approaches for counterterrorism and CVE work are using AI. To date, GNET has produced more than 20 insights related to AI and TVE trends in the last two years, which can be found at: https://gnet-research.org/resources/insights/.

If you are interested in learning more about GIFCT and its initiatives, please visit https://gifct.org/ or email us at outreach@gifct.org.

GIFCT would like to thank the Konrad-Adenauer-Stiftung for supporting this project.

7 References to the published outputs from these groups can be found in Appendix 2.

Notes

Notes

Notes

GIFCT is a 501(c)(3) non-profit organization and tech-led initiative with over 30 member tech companies offering unique settings for diverse stakeholders to identify and solve the most complex global challenges at the intersection of terrorism and technology. GIFCT's mission is to prevent TVE from exploiting digital platforms through our vision of a world in which the technology sector marshals its collective creativity and capacity to render TVE ineffective online. In every aspect of our work, we aim to be transparent, inclusive, and respectful of the fundamental and universal human rights that TVE seek to undermine.

www.gifct.org **outreach@gifct.org**