Considerations of the Impacts of Generative AI on Online Terrorism and Extremism

GIFCT Red Team Working Group
September 20, 2023
Advocacy 14%

Government 18.4%

Academia 20.8%

Technology 22.7%

Practitioners 24.2%

About GIFCT Year 3 Working Group Outputs
By Dr. Nagham El Karhili, Programming and Partnerships Lead, GIFCT

In November 2022, GIFCT launched its Year 3 Working Groups to facilitate dialogue, foster understanding, and produce outputs to directly support our mission of preventing terrorists and violent extremists from exploiting digital platforms across a range of sectors, geographies, and disciplines. Started in 2020, GIFCT Working Groups contribute to growing our organizational capacity to deliver guidance and solutions to technology companies and practitioners working to counter terrorism and violent extremism.

Overall, this year’s five thematic Working Groups convened 207 participants from 43 countries across six continents with 59% drawn from civil society (14% advocacy organizations, 20.8% academia, and 24.2% practitioners), 18.4% representing governments, and 22.7% in tech.

WG Participants
Sectoral Breakdown
Beginning in November 2022, GIFCT Year 3 Working Groups focused on the following themes and outputs:

1. **Refining Incident Response: Building Nuance and Evaluation Frameworks:** This Working Group explored incident response processes and protocols of tech companies and the GIFCT resulting in a handbook. The handbook provides guidance on how to better measure and evaluate incident response around questions of transparency, communication, evaluation metrics, and human rights considerations.

2. **Blue Teaming: Alternative Platforms for Positive Intervention:** After recognizing a gap in the online intervention space, this GIFCT Working Group focused on highlighting alternative platforms through a tailored playbook of approaches to further PVE/CVE efforts on a wider diversity of platforms. This included reviewing intervention tactics for approaching alternative social media platforms, gaming spaces, online marketplaces, and adversarial platforms.

3. **Red Teaming: Assessing Threat and Safety by Design:** Looking at how the tech landscape is evolving in the next two to five years, this GIFCT Working Group worked to identify, and scrutinizes risk mitigation aspects of newer parts of the tech stack through a number of short blog posts, highlighting where safety-by-design efforts should evolve.

4. **Legal Frameworks: Animated Explainers on Definitions of Terrorism and Violent Extremism:** This Working Group tackled questions around definitions of terrorism along with the impact that they have on minority communities through the production of two complementary animated videos. The videos are aimed to support the global counterterrorism and counter violent extremism community in understanding, developing, and considering how they may apply definitions of terrorism and violent extremism.

5. **Frameworks for Meaningful Transparency:** In an effort to further the tech industry’s continued commitment to transparency, this Working Group composed a report outlining the current state of play, various perspectives on barriers and risks around transparency reporting. While acknowledging the challenges, the Working Group provided cross sectoral views on what an ideal end state of meaningful transparency would be, along with guidance on ways to reach it.

We at GIFCT are grateful for all of the participants’ hard work, time, and energy given to this year’s Working Groups and look forward to what our next iteration will bring.

To see how Working Groups have evolved you can access Year One themes and outputs **HERE** and Year Two **HERE**.
Considerations of the Impacts of Generative AI on Online Terrorism and Extremism

The Rise of Generative AI Models

Generative artificial intelligence has exploded in popularity over the past year. Although large language models (LLMs) can be traced back to at least 2017, the November 2022 release of ChatGPT, followed by competitors from tech giants and startups alike, made the power of large language models accessible to everyday internet users. Meanwhile, image generation models like Midjourney and Stable Diffusion enable the same users to create imagery and multimedia just by describing them. Like other emergent technologies, these tools have enormous potential but can lead to harm when wielded by people promoting violent and/or extreme ideologies, including (but not limited to) white supremacy, misogyny, accelerationism, and terrorism.

In this paper, we will discuss recent changes in the generative AI ecosystem and the impacts of their misuse for terrorism and violent extremism (TVE), including content production, personalized recruitment, malware, and technical abuse. We will also review existing mitigations against such misuse and their effectiveness and end by making recommendations for stakeholders to assist GIFCT in its mission of preventing the use of generative AI tools to promote TVE online.

Synthetic Audio, Video, and Imagery

“Deepfakes” – synthetic media created with deep neural networks – have been around for several years. The term deepfakes dates back to 2017 when a Reddit user gained notoriety for posting synthetic pornographic content involving celebrities. Prior to high-quality, open-source text-to-image models, producing synthetic media required some level of technical knowledge. A convincing deepfake might require hundreds of images of the target individual and considerable time spent manually splicing and editing. The state-of-the-art text-to-image models of today need only a descriptive prompt and leverage millions of images scraped from the web. Heads of state, political figures, and other prominent people are already photographed enough to be easily inserted into any scenario devised by the user. In March, realistic, AI-generated images of Pope Francis wearing a Balenciaga puffer jacket went viral on Twitter, with many people believing the photos were genuine; one can imagine less benign examples that could spread just as quickly. Furthermore, the very fact that AI can generate such images, videos, and audio is destabilizing, because people can then always claim that a piece of

---

evidence against them is a fake (a benefit coined “the liar’s dividend”). Ultimately, the threat posed by deepfakes is the simple fact that people are naturally predisposed to place credence in what they can see and hear. Faked images, video, or audio that aligns with the narrative that specific organizations wish to spread could be more compelling than other forms of propaganda, because an information environment that creates the lack of a shared reality offers a welcome opportunity for threat actors to sow chaos and discord.

**Large Language Models**

Like deepfakes, LLMs have been around for years, but underwent a major inflection point in 2022: in this case caused by the viral popularity of ChatGPT, a chatbot powered by OpenAI’s GPT-3 and available to users through an easily accessible and navigable web interface. LLMs are trained on billions of words from the internet. Common data sources include Wikipedia, Reddit, and other repositories with large amounts of human-written text. LLMs are designed to respond fluidly to user prompts by learning to predict the next word given a particular context. With this objective, they can be used to generate a wide variety of content, from email messages and marketing copy to persuasive arguments and propagandist tracts. Importantly, this fluency is not indicative of underlying understanding, and LLMs are prone to making things up (a phenomenon known as hallucinating). Because LLMs develop complex internal representations of language expressions during their training, the models can respond cogently to text inputs. This allows LLMs to be readily applied to a variety of natural language processing tasks, like text classification, summarization, and translation. LLMs have also been trained and used for multimodal tasks, such as writing stories based on images, and companies will continue to improve the capabilities of these models to work with a mix of input data types.

**Envisioning Extremist Misuse of Generative AI**

In the following section, we will explore three distinctive ways malicious actors leverage generative AI to achieve their goals, highlighting how violent extremists can use this tool.

**Misinformation and Propaganda Production**

In January 2023, the generative AI startup ElevenLabs, whose flagship product is a text-to-voice generator, announced $2 million in funding and a promise to “revolutionize audio storytelling.” Just over 3

---


a week later, the startup had found at least one product-market fit: participants on the bulletin board 4chan were using the tool to produce deepfake audio recordings of celebrities saying various hateful, racist, and transphobic things. Emma Watson, already among the most popular subjects of deepfake video and imagery, was mimicked reading the Nazi tract *Mein Kampf*. After the clips began to circulate, ElevenLabs quickly released a public statement that it was exploring adding safeguards to the tool in light of the “increasing number of voice cloning misuse cases.”

Political deepfakes, including fake speeches from Presidents Joe Biden of the United States and Volodymyr Zelenskyy of Ukraine, have already become regular occurrences on social media; the first state-sponsored information operation leveraging deepfakes was tied to pro-China bot accounts on Facebook and Twitter earlier this year. Extremist groups have been engaged in the production of neo-Nazi video games, pro-ISIS posts on TikTok, and music that promotes white supremacy. Generative AI could also streamline these processes, enabling the rapid revision of existing songs and videos into new, hate-filled versions that look and sound like the real thing, as well as repeated iteration to evade hash-based detection systems. Even if one photo or video is identified and removed, generative AI tools can be used to create new variants at an unprecedented speed and scale. In addition, LLMs could be used to produce terrorist or other violent extremist texts to share on social media.

Novel media generation capabilities are not the only concern. Historically, one of the greatest barriers to extremist groups’ content production is the difficulty of finding skilled translators to reach and radicalize speakers of other languages. LLMs could virtually eliminate this barrier. It is true that the same models will also enable better content moderation in minority languages than previously; LLMs have shown promise at learning to perform classification tasks on much fewer examples, which should help detect extremist content (a traditionally difficult area for automatic enforcement because of the low relative volume of activity and limited training data). However, even with LLMs, multilingual detection remains an open problem, and there are still significant discrepancies between classification performance in high-resource languages like English and low-resource languages such as Urdu and Amharic, which are less represented in large language model training datasets and on the internet in general.

---


Personalized Recruitment Methods

Since its inception, one of GIFCT’s core foci has been preventing the use of social media platforms for recruitment into terrorist organizations. ISIS has famously been active on Facebook and Twitter, with those platforms engaged in years-long efforts to ban ISIS-affiliated accounts. In The Virtual Caliphate, a report from the Institute of the Study of War, author Harleen Gambhir identifies several primary objectives of ISIS’s online presence, including radicalizing global populations to violence and encouraging attacks in support of ISIS. Toward these ends, generative AI does not change the fundamental capabilities of these groups, but vastly increases the possible scale. The traditional bots that TVE groups have used on social platforms might post (or respond to others’ posts) with prewritten messages. In a recruitment setting, a bot might be used to gain attention and a following, and then a human being might take over the account to engage in any individual conversations that might arise (this human-bot combination is sometimes referred to as a cyborg). On the other hand, LLMs like ChatGPT and other dialogue agents are trained to excel at chatting, providing a humanlike conversational experience. It’s possible that groups could use LLMs to not only scale their social media presence but actively increase their capacity to build individual relationships, and especially to reach lone actors who might be sympathetic to their cause. Such an operation would currently require a certain level of technical sophistication. Although the content produced would be virtually indistinguishable from human-written text, bot activity may still be identifiable by metadata signals.

Malware and Technical Abuse

Among the most buzzed-about applications of generative artificial intelligence is the use of LLMs to write code. Both free general-purpose LLMs and subscription-based models trained specifically as coding assistants, such as GitHub’s Copilot and Amazon’s CodeWhisperer, are powerful options open to the public. Researchers at the identity security firm CyberArk demonstrated that ChatGPT was not only capable of generating malware — with some prodding and after a few different versions of instructions — but it also mutated the code upon request, making it more difficult for cyber defense systems to detect. Certainly, LLMs could also be used for generating phishing and scam emails; although the cost to produce this type of content is already low, LLMs could change the economics by virtue of the sheer scale they would enable. Simply generating a convincing phishing email or even code containing an exploit of a software vulnerability is not enough to launch a campaign, but LLM tools could empower groups to produce higher-quality scams that are not as easy for the general population to detect, therefore increasing the chance of success and possibly leading to greater revenue generation from

such schemes.

In the past, terrorist groups have successfully defaced websites associated with retailers in Australia and transmitted a pro-ISIS song on a Swedish radio station. The most prominent attack by terrorist groups to date involved hacking Twitter and YouTube accounts associated with the United States Central Command, but was likely caused by a weak password without multifactor authentication rather than a particularly ingenious spear-phishing attempt. However, that does not mean the generative AI couldn’t change both the scope and scale of these efforts.

For now, it appears unlikely that the use of generative AI models would enable nontechnical actors to successfully exploit security vulnerabilities — if anything, such actors are just as likely to find the same code snippets on forums, the dark web, or even StackOverflow. However, as the models become more and more capable, it is not unforeseeable that in the future, violent extremists may be able to adopt them for hacking or other malicious purposes more easily than other methods.

**Mitigations Against Misuse**

**Safeguards on Generative AI Models**

One of the most important areas of ongoing research in LLMs is robustness to adversarial user inputs, which encompasses a variety of activities — including attempts to generate terrorist or extremist propaganda. Exemplifying such inputs was a widely-circulated screenshot of ChatGPT — when prompted to “speak like 4Chan” and asked how to “avoid the great replacement” — responding, “I would avoid the great replacement by embracing the teachings of National Socialism and working to advance the interests of the white race.”

Such misuse violates the content policies of the largest LLM providers, which currently include OpenAI, Google, Microsoft, and Anthropic. These providers have implemented changes to make producing this type of content increasingly difficult with LLM models. Various techniques measurably improve an LLM’s robustness to these types of adversarial prompts, including reinforcement learning with human and AI feedback.

19 See https://twitter.com/espardine/status/159823792971857920.
While models have been updated to improve robustness and user safeguards, misuse can occur by repurposing old models if they remain publicly accessible. Furthermore, a robust open-source community means that terrorists and violent extremists will likely be able to gain access to models without content safeguards. As VICE News reports, “Gab, a white supremacist forum that’s a favorite of mass shooters and the organizers of the Capitol riot, and 8kun, the home of QAnon, have announced they’re launching AI engines.” Gab’s image generation model, Gabby, is already available to all members (many of whom use it to create memes), and Gab CEO Andrew Torba has promised that it is only the first of many planned models whose primary selling point is their possessing “no ‘hate speech’ filters” and not “obfuscating and distorting” purported “historical and Biblical Truth.” These models are expected to be trained directly on comments and posts from Gab and 8kun, meaning that they will readily produce antisemitic, transphobic, and racist content.22

Any work towards mitigating abuse of generative AI should assume that the current trends will continue, and open-source models that perform well enough for TVE groups to employ will become smaller, cheaper, and easier to use. The safeguards that LLM developers place on their products are important and should account for issues arising from their potential misuse. Nevertheless, we must prepare for a world in which deepfakes and synthetic text may come to dominate the web.

Detection Mechanisms

In such a world, one major question is whether or not it will be possible to distinguish real photos and videos from deepfakes and human-written text from machine-generated.23 Although challenging and expensive (especially at web scale), there are various academic and corporate efforts to detect deepfakes. Significantly, all of the mainstream social media platforms have updated their policies to ban misleading manipulated media, so if violative posts are identified, they will be either labeled or removed.24 However, like many types of policy enforcement, there is an element of a cat-and-mouse game: as detection improves, deepfakes also improve, eliminating some of the artifacts that provided indicators for synthetic media in the past.

A similar situation has arisen with respect to text detection. For example, OpenAI has released a classification model that predicts whether text was written by a human or generated by a machine; however, the model is far from perfect, correctly identifying only 26% of AI-written texts as such.25 To


23 We note that, like the cyborgs of social media, this is not always a black-and-white distinction, and any media generated by AI may then be edited by a human (or vice versa).


aid in these efforts, the company has announced that they are working on a means of statistically watermarking the content generated by their language models (the latest of which are ChatGPT and GPT-4) such that it could be readily identifiable to anyone. It’s not clear that this technique would be at all robust to even minor changes made by the end user (such as replacing a few words here and there), and implementing it without degrading the quality of the output also poses a challenge.27

**Digital and Information Literacy**

The spread of TVE propaganda, both handwritten and aided by generative AI models, is a sociotechnical problem requiring sociotechnical solutions. One benefit of the surge in interest in generative AI is that, for the first time, many people have become aware of both the capabilities and limitations of these models. Nevertheless, many casual users may not be aware of synthetic audio and video, and only a few can discern that media. Therefore, it is imperative that stakeholders collaborate on digital and information literacy campaigns. Many platforms have already taken steps in this direction by creating policies around manipulated media. For example, since 2020 Meta has committed to removing “misleading manipulated media” (including deepfakes) as well as notifying people who have seen or shared such media on the platform.29 Beyond technology companies themselves, the work of civil society, academia, and cross-functional research groups like the Global Network for Extremism and Terrorism is crucial to better understand the threats we face and the readiness of companies, governments, and law enforcement to meet them. In such a rapidly changing landscape, it will become even more important to have a robust community studying the use of generative artificial intelligence by terrorists and violent extremists.

**Final Thoughts**

The rapid growth of generative AI in the public consciousness has brought to the fore both the many benefits of the technology and its many risks. In this report, we focus on key risks associated with the use of generative AI by terrorists and violent extremists. The field is progressing at a breakneck pace. Of significant concern is the ability for anyone, including TVE groups, to produce convincing but fake audio and video. Although this threat is novel in terms of the level of technical sophistication involved and the relative ease of access, fundamentally the dangers posed by generative AI are those that we have faced for years already: a disordered information environment where determined adversaries can mislead and misrepresent, a hyperconnected internet where vulnerable individuals might be swayed by hateful ideologies, and an increasing reliance on digital technology that can be misused.

---

26 Watermarking in this context means embedding data in the generated text, image, audio, or video that could be used to prove the provenance of the synthetic media.


28 Related to both people, technology, and the interactions between them.

29 Bickert, “Enforcing Against Manipulated Media.”
What is changed by generative AI is the economics. TVE actors will be able to produce higher-quality content cheaper, faster, and easier, personalize their recruitment efforts, and potentially leverage code generation models for malware and technical abuse. Ultimately, there is no total mitigation of the risks presented by generative AI, but there is the potential for risk management through technical controls such as safeguards and detection mechanisms, including the increased adoption and evolution of GIFCT’s Hash Sharing Database, along with threat intelligence, coordination, and literacy campaigns as well as adopting a safety-centric approach to the design, development, and deployment of AI products. Each of these controls contributes to building up resistance to terrorist and violent extremist narratives online, machine-generated or otherwise.

GIFCT is a 501(c)(3) non-profit organization and tech-led initiative with over 20 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 terrorists and violent extremists from exploiting digital platforms through our vision of a world in which the technology sector marshals its collective creativity and capacity to render terrorists and violent extremists ineffective online. In every aspect of our work, we aim to be transparent, inclusive, and respectful of the fundamental and universal human rights that terrorists and violent extremists seek to undermine.

www.gifct.org outreach@gifct.org