tech against terrorism

Disrupting Terrorists Online

Early terrorist experimentation with generative artificial intelligence services

Briefing - November 2023

tech against terrorism

Disrupting terrorist activity online

Executive Summary

As a result of Tech Against Terrorism's OSINT operations, we have archived more than 5,000 pieces of AI-generated content produced by terrorist and violent extremist actors (TVEs). In the course of our work, we have found TVEs engaging with generative AI to augment current practices of creating and disseminating TVE propaganda across both Islamist and far-right ideologies.

However, this constitutes only a small fraction of the total volume of TVE content that we identify every year. We have found little evidence that generative artificial intelligence (AI) services are being systematically exploited by TVEs, and the engagement with generative AI is likely to be in its experimental phase. Our investigations indicate that there is a low risk of widespread adoption at the moment. However these experiments do indicate an emerging threat of TVE exploitation of generative AI, in the medium to long term, for the purposes of producing, adapting, and disseminating propaganda. Technical solutions, cross-industry consensus, and comprehensive policy-making are required for effective mitigation of this risk.

These instances of experimentation, by networks affiliated with Islamic State (IS), supporters of Al-Qaeda, and neo-Nazis, are summarised below:

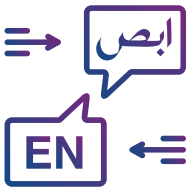
- A pro-Islamic State (IS) tech support group published a guide on 17 August advising IS networks on the secure use of an AI content generator.
- Tech Against Terrorism identified a messaging channel in early August dedicated to sharing racist, antisemitic, and pro-Nazi images which were reportedly generated using an AI art generator available on a mainstream app store.
- A "guide to memetic warfare" in which far-right propagandists are advised on using AI-generated image tools to create extremist memes.
- An IS supporter posting on an archiving platform claimed to have used an AI transcription tool to transcribe and translate a leadership message published by IS central in August 2023.
- A pro-al-Qaeda outlet has published several posters with images highly likely to have been created using a generative AI platform.
- The likely use of AI-generated imagery by actors involved in or commenting on the conflict between Israel and Hamas, following the latter's terrorist attack on southern Israel in early October 2023.

Terrorist and violent extremist (TVE) applications of generative AI

Tech Against Terrorism has developed a taxonomy that categorizes the risks posed by TVE exploitation of generative AI. The aim of this taxonomy is to provide a structured understanding of the ways in which generative AI technologies could be abused by TVE actors. Additionally, we intend to implement a "red-teaming" process to rapidly identify risk vectors that could be exploited by TVE actors. By doing so, generative AI services will receive early indicators of how their tools could be exploited by TVE early adopters. This information will enable proactive policy-making and the development of countermeasures.

Media spawning

Starting with a single image or video, a TVE actor could generate thousands of manipulated variants capable of circumventing hash-matching and automated detection mechanisms.



Automated multilingual translation

Following publication, TVE actors could translate text-based propaganda into multiple languages which would overwhelm linguistic detection mechanisms operated manually.

Fully synthetic propaganda

TVE actors could generate completely artificial TVE content. This could include speeches, images, and even interactive environments, and could overwhelm ongoing moderation efforts.

Variant recycling

TVE actors could repurpose old propaganda using generative AI tools to create "new" versions which would evade mechanisms for the hash-based detection of the original propaganda content.

Personalised propaganda

TVE actors could use AI tools to customise messaging and media to scale up the targeted recruitment of specific demographics.

Subverting moderation

TVE actors could leverage AI tools to design variants of propaganda specifically engineered to bypass existing moderation techniques.

These applications could render existing detection techniques obsolete, including database-driven approaches such as hash-matching. While generative AI poses significant risks if exploited by TVE actors, it also provides opportunities to stay ahead of the threat. Cooperation and innovation - such as collaborative red-teaming efforts – will help further understand the vulnerabilities in generative AI services and provide proactive solutions to mitigate these threats.

Extreme right-wing use of generative AI services

AI-generated images shared in dedicated neo-Nazi Telegram channel

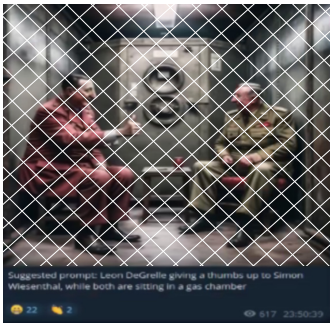
On 21 July 2023, OSINT analysts at Tech Against Terrorism identified a channel on a messaging app dedicated to sharing neo-Nazi, antisemitic, and racist images purportedly generated by artificial intelligence (AI). The channel was first active on 2 July 2023, and acts as a location for the channel administrator to post what they describe as "AI Art shitposts".

The channel contains 360 messages, the vast majority of which are images the channel administrator claims were generated using an app (see Figure 1, right). Figure 1 refers to comedian Sam Hyde. Hyde is a popular target of jokes and memes among far-right communities online who deliberately misattribute to him mass shootings and other incidents of violence, often in their immediate aftermath.

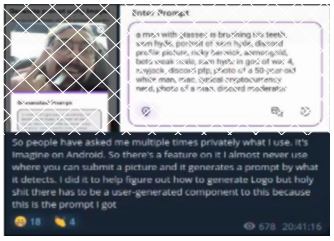
This report highlights a selection of images posted in the channel, which include what the author claims were the "suggested prompts". Many of the prompts contain overtly racist, antisemitic, or neo-Nazi themes. Many images appear to glorify historic Nazi figures, such as Adolf Hitler and Joseph Goebbels. The left image in Figure 1 displays an image of two individuals purportedly sitting in a gas chamber. According to the prompt, they are Leon Degrelle, a Waffen SS officer, and the Holocaust survivor Simon Wiesenthal.

Other posts included generated images of Hitler statues in different art styles (see middle Figure 2), overtly antisemitic posts depicting Jewish individuals as "Goblin looking" (see right Figure 2), and an image depicting Napoleon "firing into a crowd of Black Lives Matter protestors" (see left Figure 2). The app is available for Android devices and is hosted on the Google Play Store. The app's developer is reportedly based in Pakistan. [1]

Figure 1

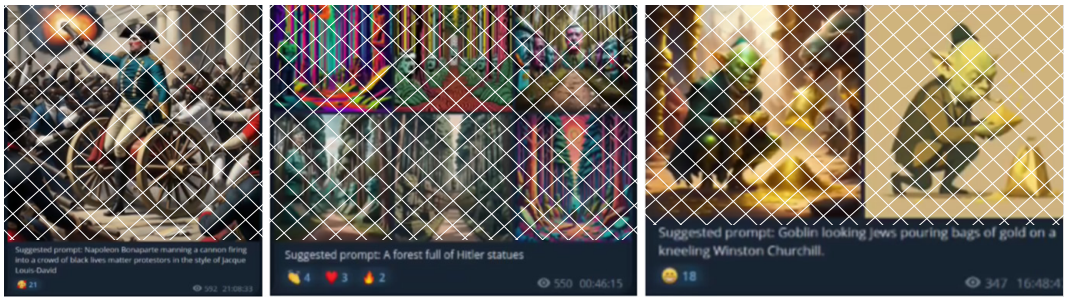


Historical figures sitting in a gas chamber



Post referring to Imagine App.

Figure 2



From left to right: racist image depicting violence against Black Lives Matter protesters; images of Adolf Hitler; antisemitic caricature.

What it means

The images generated, and the channel dedicated to propagating it, illustrate the risks associated with TVE exploitation of generative AI- and emerging AI-focused technologies. Without comprehensive and proactive policies, and technical solutions designed to counter these risks, it is likely that similar exploitation will occur across generative AI services in the medium-term future (3-12 months).

While we cannot confirm whether the channel administrator used the prompts they included in their posts, the developer's Terms and Conditions for the app state that it retains the right to remove "objectionable content." This includes "content that is hateful or advocates hate crimes".

A significant proportion of posts made in the messaging channel incite violence and hatred against religious and ethnic groups, including Jewish and Black communities. We cannot confirm at the time of writing whether the developer is moderating content generated on Imagine. However, the posts made in the channel constitute significant exploitation of the app.

1 hxxps://www.vyro[.]ai/

Guide for “AI Meme War” shared on messaging app and message board

We detected a post on a message board on 12 October in which an anonymous user, whose profile indicated that they were based in Australia, shared links to guides for creating memes and other propaganda using AI image generation tools. The post provided detailed guidance for those “making propaganda for fun,” adding that “most people” were using one AI image generator, although some were using a competitor. The guide pointed users to the image creation page of a search engine, and to a third-party editing platform to “hide [George] Floyds, [happy] merchants and other fun things in plain sight.” Imagery depicting George Floyd and the antisemitic “happy merchant” cartoon are both used widely in far-right extremist messaging.

The guides made the particular suggestion that creators should combine the image generation tools with human editing to maximise the effect of the imagery and to evade blocks on specific prompts.

What it means

The existence of guides such as the ones summarised above indicate the growing popularity of AI-generated imagery tools by propagandists promoting far-right extremist ideologies.

They also indicate that hostile users are already aware of and mitigating content moderation practices by these platforms, such as by editing imagery after it has been generated.

Figure 3

“Good vs. Bad Memes” guide, sourced from the message board.

Violent Islamist terrorist use of generative AI services

Pro-IS tech support group shares security guide for an AI content generator

On 15 August, a pro-IS tech support group shared an Arabic-language guide titled “How to protect your privacy when using [the AI content generator].” This outlet periodically produces such guides relating to different platforms, typically geared towards services widely used by IS networks. It also regularly reports on data security-related topics, such as data breaches involving tech companies, and advises on secure services such as VPNs and cryptocurrencies.

The guide offered “tips to protect your data and maintain privacy” whilst using the content generator. These tips comprised basic advice on how to disable chat history, delete chat conversations, and avoid providing “sensitive information” whilst engaging with the chat bot. It also recommended using “data anonymisation techniques”, without providing further details.

What it means

The publication of this article does not constitute proof that IS networks are widely exploiting the AI content generator for terrorist purposes, nor does the content of the article provide detailed guidance on how the tool could be used to further their cause.

However, since the topics and focus of articles produced by the outlet typically respond to the needs of IS supporters, and the platforms they are exploiting to further their cause, it is likely that IS networks are at least tentatively experimenting with the technology.

Figure 4

Guide on using the AI content generator

Pro-Islamic State user transcribes leadership message using AI transcription service

On 7 August 2023, a pro-Islamic State (IS) user of an archiving service claimed to have transcribed an Arabic-language IS propaganda message using an AI-based automatic speech recognition (ASR) system. The material was posted on 4 August on an archiving platform that is popular with IS users to host and share propaganda.

Figure 4 displays the text posted alongside the files containing the transcription: languages include Arabic, Indonesian, and English. Upon investigation of the files attached to the post, the transcription appears to have been initially done from Arabic speech to Arabic script. The insert on the bottom right displays the transcription, which includes timestamps.

The post contained a transcription of the recent statement made by al-Furqan Foundation, an official IS media outlet that in recent years has specialised in messages from the group's core leadership. The announcement declares the death of IS's previous leader, Abu Husein al-Husseini al-Qurashi, and names Abu Hafs al-Hashimi al-Qurashi as his successor.

Figure 5

Indonesian, Arabic and English-language text included in the transcription post (insert bottom right) the transcription, including timestamps.

What it means

This is the first instance in which Tech Against Terrorism has identified terrorists exploiting AI-based transcription tools for the purpose of propaganda dissemination. The task of translation and transcription is likely a time-intensive one for terrorist entities, and it is often carried out by affiliated or unofficial entities dedicated to disseminating translations of official propaganda. In several instances in recent years we have seen advertisements for translators by propaganda supporter networks.

While we have not yet identified other explicit examples of this practice, the use of such tools will, if unmitigated, significantly ease the translation and transcription process, thereby enabling terrorist propagandists to reach a broader international audience, including countries and demographics they could not previously reach. AI tools and organisations must remain vigilant against the threats of terrorist exploitation of their services and develop procedures to mitigate any attempts to misuse their tools. The transcription service in question is open source, meaning its code is publicly available.[2] While we cannot confirm how the user deployed the tool, it is likely that they used it on their own systems.

It is critical that AI developers and organisations remain aware of the emerging risks posed by TVE exploitation. As stated above, transcription and translation are time-consuming processes for terrorist entities. If AI services make these tasks easier, they are highly likely to be targeted for future exploitation. Robust information- and knowledge-sharing processes – including the sharing of the key terrorist content classifiers and indicators – could be used to train AI systems to detect abuse and prevent exploitation.

2 hxxps://github[.]com/openai/whisper

Al-Qaeda supporter outlets share propaganda highly likely generated using AI

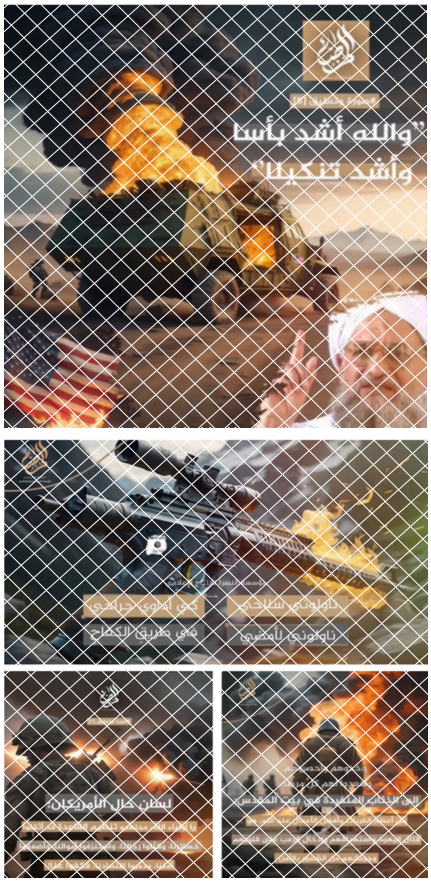
On 19 June 2023, Tech Against Terrorism OSINT analysts identified an al-Qaeda (AQ)-aligned media entity using what appears to be generative AI for propaganda production. The posts were made on a social media platform popular with violent Islamist outlets – and posted by a pro-AQ media outlet. Four instances of this have been detected since June 2023; all images were used as "posters", with quotes overlaid.

Figure 5 displays the images. There are common indicators throughout all images that suggest they are likely to have been AI-generated, such as inconsistencies in the weapon systems depicted and anomalous distortions in equipment and clothing.

There are various examples in the bottom middle image that indicate a high likelihood of AI generation. The barrel of the weapon system is bent, as are the stripes on the US flag on the soldier's left arm. The top right image includes more notable indicators that the image was generated by AI, such as the misaligned mounted scope, and what appears to be a generated barrel, handguard, and front sight at the rear of the weapon system. These inconsistencies and errors are not indicative of a stylistic approach but instead suggest AI generation. Additionally, violent Islamist imagery appears to have been superimposed onto the receiver of the weapon system.

The main left image in Figure 5 displays a burning military-style vehicle, with American flags and an image of now-deceased former al-Qaeda leader Ayman al-Zawahiri superimposed onto the image. It is highly likely that the images were not generated with a prompt including Zawahiri and the flags, but were instead incorporated into them after generation. Inconsistencies exist in this image too: differing wheel sizes and the distorted shape of the invidivual to the left of the vehicle indicate a very strong likelihood of AI generation.

Figure 6



Propaganda posters produced by the pro-Al-Qaeda outlet in June 2023.

What it means

At the time of writing, Tech Against Terrorism was unable to identify the tool(s) used to generate the images. It is likely that the users exploited free tools to mitigate the risk of identification via payment for paid services. It is highly likely that the posters were created without the text and superimposed imagery, meaning the users are not likely to have violated any terms of service relating to depictions of violence or incitement to violence and hatred. This is likely to be a significant challenge for the content moderation teams of AI tools in the event of future TVE exploitation.

It is likely that terrorist and violent extremist entities will continue to experiment with AI image generation tools in the medium-term future (3-12 months) in order to supplant and enhance ongoing propaganda creation strategies.

It is realistically possible that, more so than official media entities, unofficial violent Islamist propaganda outlets will be incentivised to exploit these tools, particularly since unofficial outlets generally lack original material with which to create propaganda. Unofficial entities are more likely to seek to stamp markers of their identity onto translated and regurgitated works, and they are also likely to be less well-resourced than their official counterparts. Because of this, it is likely that they will seek to simplify the creative processes involved in producing and promoting propaganda.

AI-Generated imagery appears online in relation to Hamas attack and ensuing conflict

Since the terrorist attack on Israel by Hamas on 7 October 2023, we have been monitoring for the appearance of AI-generated imagery online in relation to the ensuing and ongoing conflict. Our investigations show that AI-generated imagery comprises a very small proportion of online content online relating to the attack and the subsequent conflict. However, we have identified several examples of imagery relating to the conflict that is likely to have been generated or enhanced using AI tools, and in particular by Izzd Ad-din Al-Qassam Brigades. These examples are detailed below.

Several propaganda posters shared on official Al-Qassam Brigades channels since the attack are likely, in our assessment, to be AI-generated or enhanced. These posters include imagery of the group's fighters and of Israeli military targets being attacked. Examples of these are included in Figure 7 below.

We have also identified multiple examples of antisemitic AI-generated imagery shared by the English-speaking far-right. Examples of such antisemitic memes and other images are included in Figure 8 below. We are also aware of other examples of AI-generated imagery such as those detailed in the *New York Times and Wired*. [3]

Figure 7

Imagery shared by IQB on a messaging app since 7 October.

Figure 8

Anti-Semitic imagery created using Gen AI tools since the Hamas attack on Israel, sourced from a messaging board.

What it means

The above examples demonstrate the use of AI imagery during an unfolding crisis. Such imagery is likely to augment the narrative appeal, visual quality, and aesthetic professionalism of content produced by malicious actors and thereby increase the scope for mis- and disinformation attendant on such events. However, in our view, the prevalence of AI-generated imagery online in relation to the Israel-Hamas war is negligible compared to authentic content.

3 https://www.wired.co.uk/article/israel-hamas-war-generative-artificial-intelligence-disinformation, https://www.nytimes.com/2023/10/28/business/media/ai-muddies-israel-hamas-war-in-unexpected-way.html