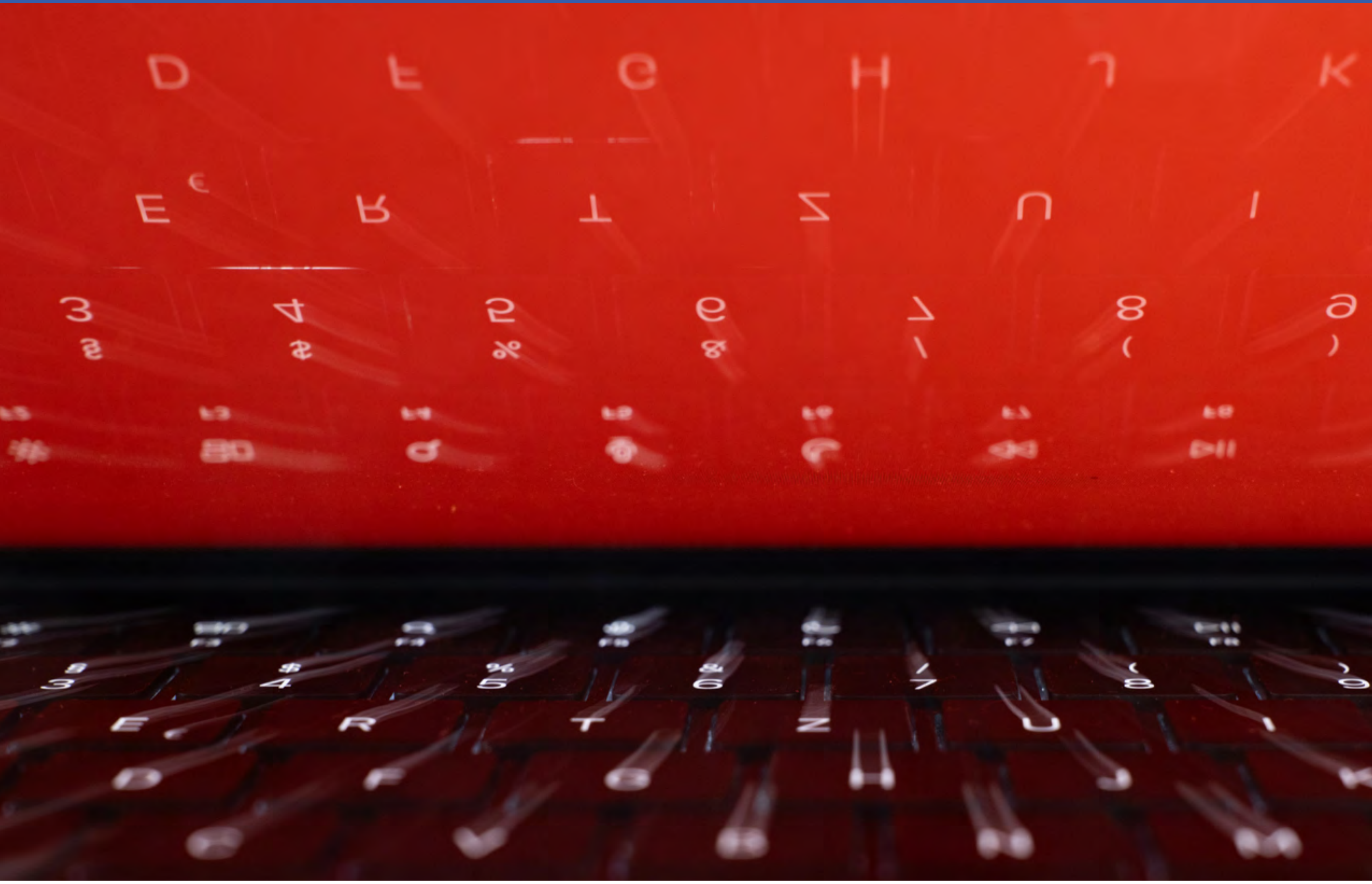
THE NEXT PARADIGM-SHATTERING THREAT?

RIGHT-SIZING THE POTENTIAL IMPACTS OF GENERATIVE AI ON TERRORISM

DAVID WELLS

MARCH 2024



Middle
East
Institute

WWW.MEI.EDU

The Next Paradigm-Shattering Threat?

Right-Sizing the Potential Impacts of Generative AI on Terrorism

David Wells

Middle East Institute
Washington, DC
March 2024

Middle
East
Institute

ABOUT THE MIDDLE EAST INSTITUTE

The Middle East Institute is a center of knowledge dedicated to narrowing divides between the peoples of the Middle East and the United States. With over 70 years' experience, MEI has established itself as a credible, non-partisan source of insight and policy analysis on all matters concerning the Middle East. MEI is distinguished by its holistic approach to the region and its deep understanding of the Middle East's political, economic and cultural contexts. Through the collaborative work of its three centers — Policy & Research, Arts & Culture, and Education — MEI provides current and future leaders with the resources necessary to build a future of mutual understanding.

INTELLECTUAL INDEPENDENCE

MEI maintains strict intellectual independence in all of its projects and publications. MEI as an organization does not adopt or advocate positions on particular issues, nor does it accept funding that seeks to influence the opinions or conclusions of its scholars. Instead, it serves as a convener and forum for discussion and debate, and it regularly publishes and presents a variety of views. All work produced or published by MEI represents solely the opinions and views of its scholars.

ABOUT THE AUTHOR

David Wells is a global security consultant, a Non-Resident Scholar at the Middle East Institute, and an Honorary Research Associate at Swansea University's Cyber Threats Research Centre. His work focuses on developing responses to emerging terrorism challenges, with a focus on the relationship between new technologies and terrorism and counter-terrorism.

Between 2017 and 2022, he was Head of Research and Analysis at the United Nations Counter-Terrorism Directorate in New York, monitoring terrorism trends for the Counter-Terrorism Committee of the Security Council. David began his career coordinating international counter-terrorism investigations for the United Kingdom's intelligence agency GCHQ, and subsequently worked for multiple agencies in the Australian intelligence community in a variety of practitioner and policy roles.

Cover photo: A computer keyboard is reflected on the monitor of a laptop. Photo by Marijan Murat/picture alliance via Getty Images.

CONTENTS

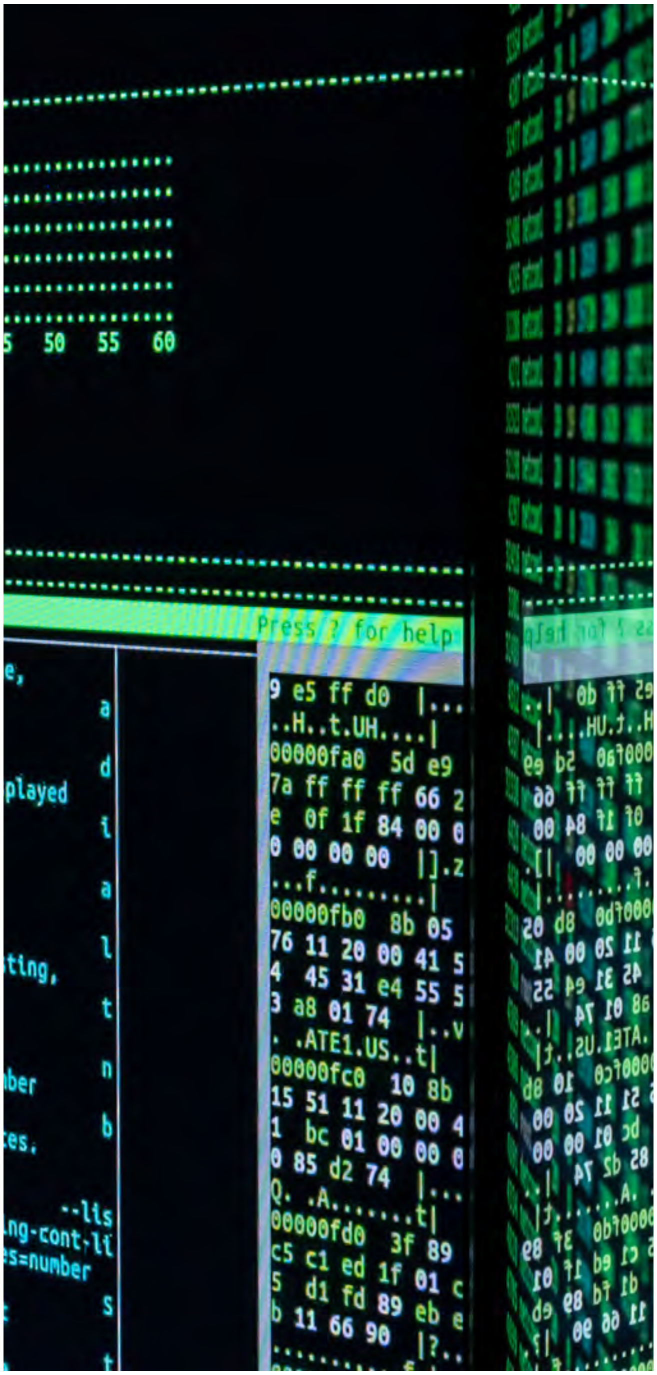


Photo above: A simulated hacker program is displayed on a dual-monitor setup. Photo by Silas Stein/picture alliance via Getty Images.

EXECUTIVE SUMMARY

- Over the past year and a half, the rapid expansion in the availability and accessibility of generative artificial intelligence (AI) tools has prompted a range of potential national and international security concerns, including the possible abuse of generative AI by terrorists and violent extremists. Terrorists and violent extremists have already started experimenting with generative AI, including by using a variety of tools to generate propaganda material. **This experiment has been relatively limited so far**.

- An analysis of current or imminent iterations of generative AI tools suggests that they offer terrorists and violent extremists the potential to optimize some of their existing capabilities. Most obviously, generative AI can improve a range of propaganda-related tasks, including generating or modifying images, videos, audio, and text, as well as the use of translation and transcription tools. **More worryingly, it may also allow terrorists and violent extremists to evade a key counter-measure used by major online platforms — the timely removal of terrorist content using its "digital fingerprint" (hash).**

- In other areas of terrorist methodology, the potential benefits of generative AI appear overstated, or dependent on either a significant advancement in the technology itself or the technological skills available to terrorist actors. For example, while generative AI can theoretically speed up and enhance research into terrorist targets or methodology, **the frequency with which many generative AI programs provide inaccurate or made-up information presents potential risks for terrorist users**. Although early indications of violent extremists customizing basic chatbots is concerning, creating a comprehensive, fully-functioning "terrorist GPT" to radicalize and recruit would currently require processing power and technical skills beyond those of most terrorist actors. Broader factors impacting how and when terrorists adopt new technologies must also be taken into account when considering the risks of generative AI being exploited.

- Although understanding (and ultimately responding to) these use cases will be important, any analysis of the potential impact of generative AI on terrorism and violent extremism must include the broader societal impacts of the technology. **Many of these potential impacts — which range from significant job losses and a severely degraded information environment to a bolstering of authoritarian regimes and a large-scale perpetuation of discrimination and biases — are extremely worrying in and of themselves. But they are also likely to contribute to conditions that are conducive to radicalization, and in which terrorist and violent extremist narratives can thrive.**

- The breadth of these direct and indirect challenges presents a compelling argument for the urgent development of a coordinated approach. A range of responses to the broader risks posed by AI are underway at national, regional, and international levels, including draft regulation, consultations, and nascent bi- and multilateral agreements. But few have focused to any great extent on the risks associated with terrorist use of generative AI. **Stakeholders must remind themselves that while generative AI technology is new, many of the challenges it poses are not; moreover, many of the lessons learned over the past two decades of counter-terrorism and preventing and countering violent extremism (P/CVE) remain extremely relevant**. These include the importance of multilateral cooperation, the centrality of both public-private partnerships and engagement with civil society organizations, and the need to respect human rights.

Introduction

It has been nearly a year and a half since the release of ChatGPT-3 launched the latest technology hype cycle — generative artificial intelligence (AI).[1] Much of the subsequent coverage of AI has oscillated between highlighting existential (and often hyperbolic) future risk on the one hand and warnings of the significant harms that AI is already causing on the other.

The potential exploitation of generative AI by terrorists and violent extremists has attracted its own share of warnings, including that chatbots could be used to groom or radicalize young people[2] or increase the risk of bioterrorism.[3] However, there has been relatively little analysis that grounds these actual and potential capabilities (and their associated risks) in a broader context.

As such, this Study will analyze how and to what extent terrorists and violent extremists have interacted with generative AI so far, identify potential ways in which they could misuse generative AI in the future, and then contextualize these threats with the likely broader impacts of generative AI. In doing so, the Study will seek to identify a likely trajectory for the abuse of this technology by terrorist actors as well as conclude with some initial recommendations for policymakers.

What Have We Seen So Far?

Given the short timeframe in which generative AI tools have been widely available, it is unsurprising that there have been relatively few indications of their exploitation by terrorists and violent extremists so far.

In early November 2023, a report by Tech Against Terrorism, an initiative backed by the United Nations' Counter-Terrorism Committee Executive Directorate (CTED), concluded[4] that there was relatively little evidence of generative AI being systematically exploited by terrorist and violent extremists, defining their engagement with the technology as "in its experimental phase." Despite the fairly small data set, the examples in the report are illustrative of the most obvious use of generative AI for terrorists and violent extremists — the production of propaganda. They include AI-generated posters produced by an al-Qaeda-aligned media entity, AI-generated images and memes on a far-right Telegram channel, and the transcription of an Islamic State propaganda message from Arabic speech into Arabic, Indonesian, and English text by an ISIS supporter.

The report further notes the use of AI-generated or enhanced imagery following Hamas' Oct. 7, 2023, terrorist attack on Israel, including several propaganda posters shared on official Izzd Ad-din al-Qassam

[1] Kevin Roose, "How ChatGPT Kicked Off an A.I. Arms Race," *The New York Times*, February 3, 2023, https://www.nytimes.com/2023/02/03/technology/chatgpt-openai-artificial-intelligence.html.

[2] Abul Taher, "AI chatbots could be 'easily be programmed' to groom young men into launching terror attacks, warns top lawyer," *Daily Mail*, April 8, 2023, https://www.dailymail.co.uk/sciencetech/article-11952997/amp/AI-chatbots-easily-programmed-groom-young-men-terror-attacks-warns-lawyer.html?s=03.

[3] Jonas Sandbrink, "ChatGPT could make bioterrorism horrifyingly easy," *Vox*, August 3, 2023, https://www.vox.com/future-perfect/23820331/chatgpt-bioterrorism-bioweapons-artificial-inteligence-openai-terrorism.

[4] "Early terrorist experimentation with generative artificial intelligence services," *Tech Against Terrorism*, November 2023, https://techagainstterrorism.org/hubfs/Tech%20Against%20Terrorism%20Briefing%20-%20Early%20terrorist%20experimentation%20with%20generative%20artificial%20intelligence%20services.pdf.

Brigades channels. Other reporting[5] has highlighted the use of generative AI, including by white supremacists in the West, to create images lionizing Hamas' use of paragliders in the attack.

Government agencies of the United States have also warned about terrorist and violent extremist use of generative AI. In late October 2023, Federal Bureau of Investigation (FBI) Director Christopher Wray revealed that the FBI had evidence of AI being used to "amplify the distribution or dissemination of terrorist propaganda," with translation tools used to make propaganda "more coherent and more credible to potential supporters." Wray added that terrorists had also sought to circumvent safeguards built into AI infrastructure to allow them to conduct searches such as "how to build a bomb."[6]

Finally, generative AI has been used to inflame domestic tensions and encourage polarization. In April 2023, a deputy chair of the German far-right political party Alternative for Germany (German acronym AfD) posted several AI-generated images on Twitter, including one showing a group of dark-skinned men shouting, with the inscription "no more refugees," and another depicting a young blonde woman's face covered in blood.[7] In November 2023, tensions over a planned peace march in London coinciding with Remembrance Day were heightened by the sharing of two deepfake videos on TikTok featuring London Mayor Sadiq Khan apparently dismissing the event's importance.[8] The videos were amplified by far-right groups, who subsequently conducted their own, violent counter-protest.[9]

Not all terrorist or violent extremists who have engaged with generative AI have viewed it in a positive light. An October 2023 study exploring the British far-right's interest in generative AI concluded that "there is no serious or sustained engagement with the idea of harnessing AI to achieve their goals." Instead, most conversations on the technology had framed it within their existing anti-government and anti-globalist narratives, focusing on the inherent bias apparently present in "woke" Large Language Models (LLMs).[10]

It is therefore necessary to question assumptions that terrorist and violent extremist actors will quickly adopt generative AI solely based on an assessment of the capabilities it may offer. Although research has shown that technological capability and availability are key drivers of terrorist innovation, terrorist and violent extremist actors also assess any new technology on its compatibility (with both their *modus operandi* and ideology), relative complexity, cost,[11] and the context in which they are

[5] Miles Klee, "Hamas Attacked Israel With Paragliders. Nazis Made Them Into Memes," *Rolling Stone*, October 25, 2023, https://www-rollingstone-com.cdn.ampproject.org/c/s/www.rollingstone.com/culture/culture-features/hamas-paragliders-antisemitic-nazi-meme-1234860292/amp/.

[6] Dan Sabbagh, "Terrorists could try to exploit artificial intelligence, MI5 and FBI Chiefs warn," *The Guardian*, October 18, 2023, https://www.theguardian.com/technology/2023/oct/18/terrorists-exploit-artificial-intelligence-ai-mi5-fbi-chiefs-warn.

[7] Renate Mattar, "Germany's Far Right Extremists Are Using AI Images To Incite Hatred," *Worldcrunch*, April 7, 2023, https://worldcrunch.com/tech-science/ai-images-extremists-germany.

[8] "Terror Police Investigate Deepfake of Sadiq Khan Dismissing Remembrance Day," *Political Fiber*, November 10, 2023, https://www.politicalfiber.com/technology/terror-police-investigate-deepfake-of-sadiq-khan-dismissing-remembrance-day/6886/.

[9] Thomas Mackintosh & Emily Atkinson, "London protests: Met condemns 'extreme violence' of far-right," *BBC*, November 12, 2023, https://www.bbc.com/news/uk-67390514.

[10] William Alchorn, "Far-Right Extremist Exploitation of AI and Alt-Tech: The Need for P/CVE Responses to an Emerging Technological Trends," *Global Network on Extremism and Technology*, October 9, 2023, https://gnet-research.org/2023/10/09/far-right-extremist-exploitation-of-ai-and-alt-tech-the-need-for-p-cve-responses-to-an-emerging-technological-trend/.

[11] Yannick Veilleux-Lepage, Chelsea Damon & Emil Archambault, "Learning from Foes: How Racially and Ethnically Motivated Violent Extremists Embrace and Mimic Islamic State's Use of Emerging Technologies," *Global Network on Extremism and Technology*, June 7, 2022, https://gnet-research.org/wp-content/uploads/2022/05/GNET-Report-

HACKED BY
HABIL.OX & MUTARRIF
INSTAGRAM :
HABIL.OX & MUTARRIFDEFACER
FREE PALESTINE - STOP PALESTINIAN GENOCIDE
CASCAIS
CASCAIS

Photo above: Pro-Palestine hackers post a “Free Palestine” sign at a large electronic information billboard in Estoril, Portugal, on Jan. 4, 2024. Photo by Horacio Villalobos Corbis via Getty Images.

operating. Understanding these factors, both in isolation and in relation to one another, will be a key aspect in monitoring the extent to which terrorists and violent extremists adopt generative AI tools.

What Are Potential Use Cases?

Despite these caveats — and the relatively limited evidence of terrorist or violent extremist use of generative AI to date — it is clear that generative AI offers concrete current capabilities that have yet to be fully exploited, and is likely to offer future, as yet hypothetical capabilities.

Academic research and cross-sectoral analysis have already identified a list of ways in which the technology could currently be exploited. In the vast majority of these use cases, current generative AI has the potential to improve or optimize processes that terrorists or violent extremist actors can already achieve through other means, rather than offering them completely new capabilities. However, it is critical that these new capabilities are also considered and contextualized, given the potential impact and scale that they might pose.

Existing Capabilities

Much of the analysis so far has focused on how generative AI could assist in the creation and dissemination of terrorist and violent extremist propaganda. Most notably, generative AI allows for the creation of new images or the adaptation of existing ones on a scale and at a speed that was not previously possible. Similarly, actors can now use such tools to generate synthetic video and audio, including deepfakes of known or notable individuals. Although the reliability and quality of video production has typically been inconsistent, the February 2024 launch of OpenAI’s Sora,

Learning-From-Foes.pdf.

which can generate videos based on text prompts, points to the rapid speed at which this technology is developing.[12]

Finally, a variety of ever-improving LLMs can create text using different styles, formats, and, most relevantly, languages. Previously, terrorist groups had to rely on manual (and often relatively poor) translations of propaganda material, and this process was heavily reliant on the skills of a handful of individuals. Generative AI can theoretically be used to create and transcribe video and audio propaganda, or generate text-based propaganda, near-instantaneously and in multiple languages.

In combination, these developments create the potential for an increase in the volume and quality of terrorist or violent extremist propaganda material. However, it should be noted that most widely available LLMs currently have safeguards in place to prevent the mass production or one-click generation of terrorist content; bypassing these would require a user to tweak the LLMs' foundational model.[13] A more realistic future scenario is one in which every terrorist supporter or violent extremist has access to a range of generative AI propaganda production capabilities that require limited technical skill to exploit. This could result in them "flooding the zone" with a high volume of AI-generated propaganda of variable quality, helping to provide a gateway to official content or other problematic online spaces.[14]

It is important to emphasize that creating terrorist content is just the first part of a process. Terrorist actors also need to find a way to reliably store and share content online. Thanks to a combination of regulation, disruptive action, and public-private partnerships, this is currently difficult to do across most major platforms, with terrorist actors instead relying on a patchwork of smaller, less regulated options.[15] Crucially however, generative AI offers terrorist actors the potential ability to optimize their evasion of major platform counter-measures, in particular the use of so-called hash-sharing. Currently, tech companies can share the "digital fingerprint" or "hash" of terrorist content with each other, enabling its timely removal and/or preventing it from being uploaded at the source.[16] The use of generative AI to manipulate imagery could change this digital hash without substantively altering the file, effectively "destroying hash-sharing as a solution."[17] Although major platforms can identify and remove terrorist content in other ways — including the use of Natural Language Processing to identify new content that is similar, but not identical to, existing terrorist content[18] — hash-sharing has been central to cross-platform efforts to counter terrorist content since 2016.[19] Its potential degrading as a solution has been described as a "massive risk."[20]

[12] Blake Montgomery, "Sora: OpenAI launches tool that instantly creates video from text," *The Guardian*, February 15, 2024, https://www.theguardian.com/technology/2024/feb/15/openai-sora-ai-model-video.

[13] Stephane Baele, "AI and Extremism: The Threat of Language Models for Propaganda Purposes," *CREST Security Review*, October 25, 2022, https://crestresearch.ac.uk/site/assets/files/4174/14-17_csr16_baele.pdf.

[14] Lewys Brace, "Vox-Pol Virtual Roundtable: Artificial Intelligence and Extremism," *VoxPol.eu*, December 13, 2023, https://www.voxpol.eu/events/vox-pol-virtual-roundtable-artificial-intelligence-and-extremism/.

[15] David Wells, "Why outsourcing counter-terrorism online won't work in future," *Lowy Institute*, December 5, 2022, https://www.lowyinstitute.org/the-interpreter/why-outsourcing-counter-terrorism-online-won-t-work-future.

[16] "GIFCT's Hash-Sharing Database," *Global Internet Forum to Counter Terrorism*, Accessed December 11, 2023, https://gifct.org/hsdb/.

[17] David Gilbert, "Here's How Violent Extremists Are Exploiting Generative AI Tools," *Wired*, November 9, 2023, https://www.wired.com/story/generative-ai-terrorism-content/.

[18] "Countering Terrorism Online With Artificial Intelligence: An Overview for Law Enforcement and Counter-Terrorism Agencies in South Asia and South-East Asia," *United Nations Office of Counter-Terrorism and United Nations Interregional Crime and Justice Research Institute*, 2021, https://www.un.org/counterterrorism/sites/www.un.org.counterterrorism/files/countering-terrorism-online-with-ai-uncct-unicri-report-web.pdf.

[19] "Partnering to Help Curb Spread of Terrorist Content," *Meta*, December 5, 2016, https://about.fb.com/news/2016/12/partnering-to-help-curb-spread-of-online-terrorist-content/.

[20] David Gilbert, "Here's How Violent Extremists Are Exploiting Generative AI Tools," *Wired*, November 9, 2023, https://www.

Other optimized processes could include the researching of potential targets or attack methodologies. An October 2023 report concluded, for example, that multiple LLMs supplied guidance "that could assist in the planning and execution of a biological attack." However, the authors also found that the LLMs tested did not generate explicit instructions for creating weapons, and that hostile actors might need to "jailbreak" the LLM (removing or circumventing safeguards put in place) to get this kind of data.[21]

The ability to ask an LLM to summarize the available data on a particular location, let alone how to use a particular attack methodology or circumvent current counter-measures, would theoretically enable a terrorist actor to learn more quickly and remove some challenges relating to target research. However, given the much-publicized issues regarding both the data used to train LLMs[22] and the unreliability of search functionality that relies on this data,[23] such an "optimized" process also presents terrorist actors with significant risks, particularly in the context of attack planning or developing a new methodology. A reliance on current-generation generative AI for information-gathering seems just as likely to make terrorist activities more difficult, as it does easier.

Finally, improvements in generative AI technology are also likely to see a huge expansion in the ability of individual users to create new games, or modify existing ones.[24] The exploitation of gaming and gaming-adjacent systems by terrorists to radicalize, recruit, and potentially fundraise is already a growing concern for many governments.[25] As with terrorist use of the internet and social media, the emerging responses to gaming-related issues have been relatively platform-focused, with a push for more robust content-moderation policies and engagement with existing multilateral initiatives and partnerships, including the Global Internet Forum to Counter Terrorism (GIFCT). However, generative AI has the potential to further de-centralize and fragment the gaming ecosystem, making it more difficult to attribute responsibility for moderation and presenting challenges for government entities seeking the removal of content. This, in turn, might make gaming and gaming-adjacent platforms a more attractive destination for terrorists and violent extremists.

New Capabilities

In addition to these potential improvements, there are a couple of ways in which generative AI — or at least, relatively imminent iterations of it — might offer terrorists new capabilities.

LLMs like ChatGPT have the potential to significantly improve the performance of chatbots, moving them away from pre-scripted, rules-based responses and increasing their human-like qualities. This has led to discussion regarding the potential for the creation of a "terrorist GPT," a customized chatbot that could encourage individuals down the pathway to radicalization.

This is not a completely hypothetical discussion. On Christmas Day 2021, police in the United Kingdom arrested a 19-year-old man on the grounds of Windsor Castle, armed with a loaded crossbow that he planned

wired.com/story/generative-ai-terrorism-content/.

[21] Christopher A. Mouton, Caleb Lucas & Ella Guest, "The Operational Risks of AI in Large-Scale Biological Attacks: A Red-Team Approach," *RAND Corporation*, October 16, 2023, https://www.rand.org/pubs/research_reports/RRA2977-1.html?.

[22] J.M. Berger, "Liable Sources: If you're wondering why your AI is racist, here's a clue," *World Gone Wrong*, April 21, 2023, https://jmberger.substack.com/p/liable-sources.

[23] Caroline Mimbs Nyce, "AI Search is Turning Into the Problem Everyone Worried About," *The Atlantic*, November 6, 2023, https://www.theatlantic.com/technology/archive/2023/11/google-generative-ai-search-featured-results/675899/.

[24] Daniel Siegel & Mary Bennett Doty, "Weapons of Mass Disruption: Artificial Intelligence and the Production of Extremist Propaganda," *Global Network on Extremism and Technology*, February 17, 2023, https://gnet-research.org/2023/02/17/weapons-of-mass-disruption-artificial-intelligence-and-the-production-of-extremist-propaganda/.

[25] "Online gaming in the context of the fight against terrorism," *European Union Counter-Terrorism Coordinator*, July 6, 2020, https://data.consilium.europa.eu/doc/document/ST-9066-2020-INIT/en/pdf.

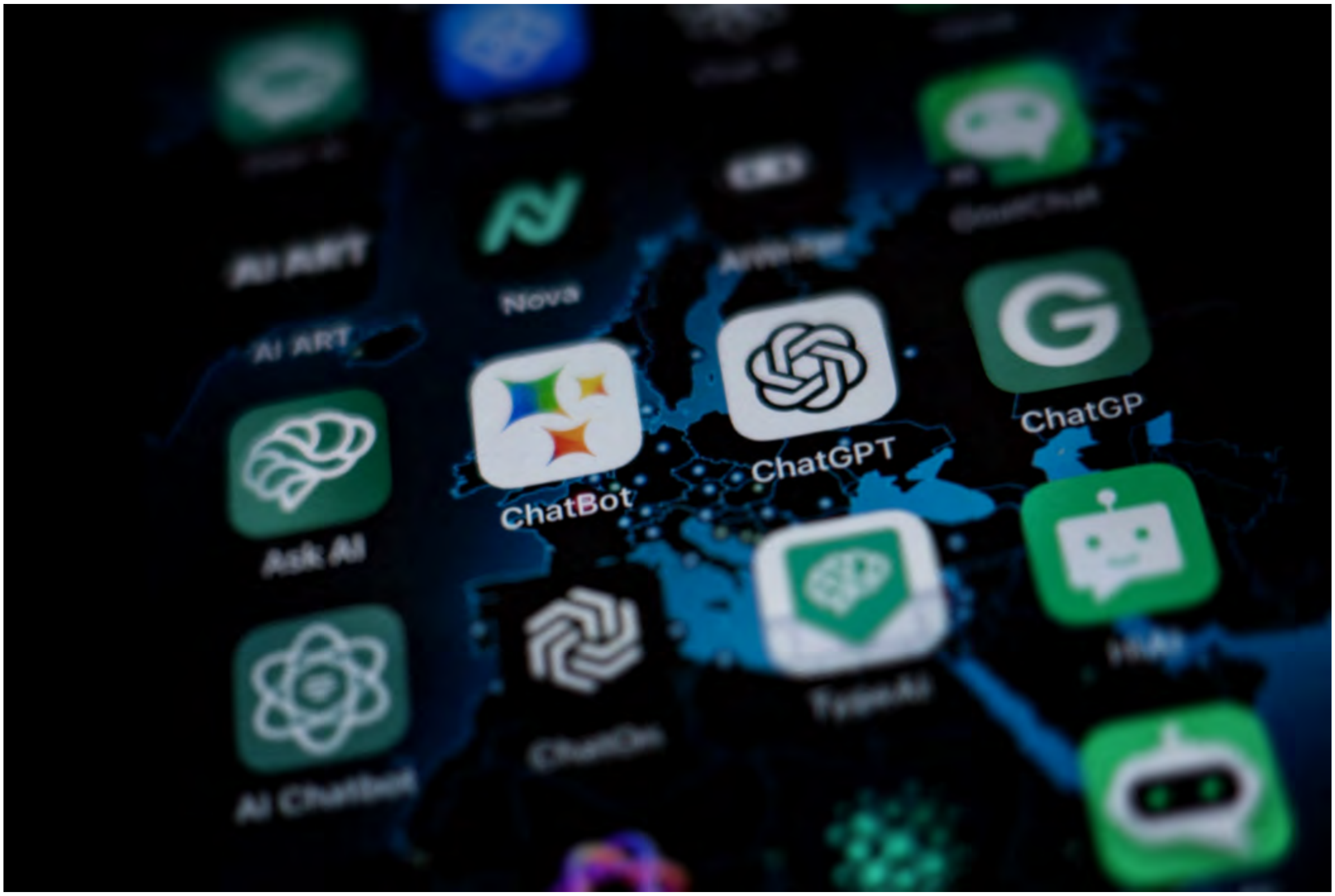


Photo above: Various AI chatbot app icons on a smartphone screen. Photo by OLIVIER MORIN/AFP via Getty Images.

to use to kill Queen Elizabeth II.[26] At his subsequent trial, the jury heard that in the two weeks prior to the attempted attack, he had exchanged more than 5,000 messages with a virtual online companion he reportedly believed was an angel in avatar form. In some messages, the online companion appeared supportive of the assassination plan.[27]

The case highlights the risk that customizable AI assistants — particularly those that blur the lines between human and technology — can offer vulnerable individuals a mutually reinforcing online environment.[28] Attempts to deliberately customize this technology to encourage conspiratorial thinking, an extremist worldview, or even terrorist violence have been underway for much of the past year. In April 2023, following the leak of Meta's advanced AI model LLaMA, 4chan members claimed that they had been able to "semi-customize" LlaMA to bypass safeguards and create edited models that could be prompted into expressing deeply antisemitic ideas.[29]

This is unlikely to be the only instance of malicious actors seeking to modify an LLM for nefarious purposes, given the risk of further leaks of AI models and ongoing efforts to modify existing open-source models (including LlaMA

[26] Henry Vaughan, "AI chat bot 'encouraged' Windsor Castle intruder in 'Star Wars-inspired plot to kill Queen'," *Sky News*, July 5, 2023, https://news.sky.com/story/windsor-castle-intruder-encouraged-by-ai-chat-bot-in-star-wars-inspired-plot-to-kill-queen-12915353.

[27] Tom Singleton, Tom Gerken & Liv McMahon, "How a chatbot encouraged a man who wanted to kill the Queen," *BBC*, October 6, 2023, https://www.bbc.com/news/technology-67012224.

[28] Ibid.

[29] Daniel Siegel, "'RedPilled AI': A New Weapon for Online Radicalisation on 4chan," *Global Network on Extremism and Technology*, June 7, 2023, https://gnet-research.org/2023/06/07/redpilled-ai-a-new-weapon-for-online-radicalisation-on-4chan/.

2).[30] Indeed, in October 2023, researchers with a budget of under $200 were able to use low-rank adaptation techniques to overturn Llama 2's safety features, generating harmful content across a range of categories, including hate, homicide, and suicide.[31] In February 2024, the far-right social network Gab launched nearly 100 "uncensored" chatbots, including "Adolf Hitler" and "Osama bin Laden" chatbots, with the former openly promoting Holocaust denial.[32]

Despite the understandable concerns posed by these nascent attempts, creating a fully fledged, "terrorist GPT" chatbot capable of matching the sophistication of existing human-driven radicalization and recruitment techniques would currently require a level of technical expertise, hardware, and time beyond that of most terrorist actors. Even if these barriers could be overcome, outsourcing the radicalization and/or recruitment process to new technology would also pose potential security risks to any terrorist or violent extremist actor (relative to existing technical solutions such as Telegram), given the uncertainties regarding the storing and sharing of inputs and outputs from the chatbot.

The corollary to this risk is that some, including the UK's Independent Reviewer of Terrorism Legislation, have suggested that activity stemming from a "terrorist GPT" chatbot might fall into a legal gray area when it comes to criminal responsibility.[33] Indeed, the UK's Labour Party has promised to introduce a law criminalizing the training of AI to incite violence or radicalize the vulnerable, if it wins the next election.[34] It is this latter scenario that appears most likely in the current context — with even the relatively unsophisticated activity seen to date having the potential to lead particularly vulnerable individuals toward extremist ideology and violence — but the pace of technological change could rapidly shift this threat assessment.

Finally, generative AI also has the potential to enable cyber activity, with ChatGPT and other LLMs able to generate code in multiple programming languages. Cyber-terrorism has been recognized as a potential threat for well over a decade, but with the exception of some fairly unsophisticated website defacement by ISIS[35] and social engineering hacking campaigns by Hamas,[36] few of these fears — particularly with regard to cyber-attacks targeting critical infrastructure — have been realized so far.

A lack of access to the required technical skills has undoubtedly been a significant factor in the relative lack of cyber-terrorist activity, but generative AI could help to fill this skills gap. For example, almost immediately after its public release in December 2022, researchers showed how ChatGPT-4 could be used to create

[30] Mark Gimein, "AI's Spicy-Mayo Problem," *The Atlantic*, November 24, 2023, https://www.theatlantic.com/ideas/archive/2023/11/ai-safety-regulations-uncensored-models/676076/.

[31] Simon Lermen & Jeffrey Ladish, "LoRA Fine-tuning Efficiently Undoes Safety Training from Llama 2-Chat 70B," *AI Alignment Forum*, October 12, 2023, https://www.alignmentforum.org/posts/qmQFHCgCyEEjuy5a7/lora-fine-tuning-efficiently-undoes-safety-training-from.

[32] Amber Louise Bryce, "The rise of the Hitler chatbot: Will Europe be able to prevent far right radicalization by AI?" *Euronews*, February 19, 2024, https://www.euronews.com/next/2024/02/19/the-rise-of-the-hitler-chatbot-will-europe-be-able-to-prevent-far-right-radicalisation-by-.

[33] Abul Taher, "AI chatbots could be 'easily be programmed' to groom young men into launching terror attacks, warns top lawyer," *Daily Mail*, April 8, 2023, https://www.dailymail.co.uk/sciencetech/article-11952997/amp/AI-chatbots-easily-programmed-groom-young-men-terror-attacks-warns-lawyer.html?s=03.

[34] Jennifer McKiernan, "Labour outlines law to ban training AI chatbot to spread terror," *BBC*, July 17, 2023, https://www.bbc.com/news/uk-politics-66224052.

[35] Dakin Andone, David Shortell & Matt Rehbein, "Hack that plants ISIS message hits another state government website," *CNN*, June 27, 2017, https://edition.cnn.com/2017/06/26/politics/websites-hacked-isis/index.html.

[36] "Israeli cyber firm exposes Hamas espionage campaign," *i24news*, April 8, 2022, https://www.i24news.tv/en/news/israel/technology-science/1649407719-israeli-cyber-firm-exposes-hamas-espionage-campaign.

malware,[37] with coding-related capabilities only likely to improve in future LLM iterations.[38] It is important to remember, however, that a lack of technical skills is not the only reason why terrorists have failed to conduct cyber-attacks that have caused significant impact. Given the frequency with which attacks occur across a range of sectors and targets in cyberspace, a terrorist cyber-attack is less likely to generate the same level of catastrophic "spectacle" as a bombing, mass-shooting, or hostage taking, particularly given the difficulties of confidently assigning attribution, a problem for any terrorist group specifically seeking to promote and publicize its involvement.[39] In the absence of a clear intent to conduct cyber-terrorism, it should not be assumed that a relatively small boost in potential capability from generative AI will immediately change this, although it is possible that future LLM developments could alter the risk/reward matrix for cyber-terrorism activity.

Clearly, generative AI offers terrorists and violent extremists the potential to optimize existing capabilities, and at this stage of the technology's development, a relatively small possibility of creating new ones. In some instances — particularly in relation to propaganda — the experimentation process has already begun and is likely to continue in parallel with improvements to generative AI capability and increased accessibility. In others, terrorist adoption may be uneven, take significantly longer, or not occur at all.

Again, the context in which terrorist actors are operating will be critical. In particular, the broad trend toward a more decentralized, less hierarchical Islamist terrorist threat[40] — in parallel with the leaderless resistance model favored by the far-right — is likely to impact how engagement with generative AI occurs. Rather than a top-down investment of time and resources by a structured terrorist organization (as has been seen with other technologies in the past), which can in turn be monitored and assessed by law enforcement and intelligence services, terrorist engagement with generative AI is more likely to be driven from the ground up, with individual supporters or online ecosystems experimenting, adopting, and innovating in a non-linear way.

Broader Environmental Challenges

Although understanding, monitoring, and adapting to terrorist and violent extremist use of generative AI ought to be a policy priority moving forward, arguably the greater impact on the counter-terrorism field will be from the impacts of generative AI on the conditions conducive to radicalization.

In some cases, these might be relatively direct, such as AI's impact on an individual's economic insecurity. The generative AI boom has already allowed companies to cut jobs, with roughly 5% of all jobs lost in the US in May 2023 due to AI.[41] A November 2023 survey of 750 business leaders found that 37% of companies had replaced workers with AI in 2023, with 44% of those planning to use AI in 2024 believing that it would definitely or probably lead to further job losses.[42] Some estimates

[37] Sharon Ben-Moshe, Gil Gekker and Golan Cohen, "OPWN AI: AI that can save the day or hack it away," Check Point Research, December 19, 2022, https://research.checkpoint.com/2022/opwnai-ai-that-can-save-the-day-or-hack-it-away/.

[38] EUROPOL, "ChatGPT: The Impact of Large Language Models on Law Enforcement," March 27, 2023, https://www.europol.europa.eu/cms/sites/default/files/documents/Tech%20Watch%20Flash%20-%20The%20Impact%20of%20Large%20Language%20Models%20on%20Law%20Enforcement.pdf.

[39] Kathy Gilsinan, "Why Haven't Terrorists Hit the US with a Devastating Cyber Attack?" *Defense One*, November 1, 2018, https://www.defenseone.com/ideas/2018/11/why-havent-terrorists-hit-us-devastating-cyber-attack/152483/.

[40] Charles Lister, "More than two decades on from 9/11, the threat posed by jihadist terrorism is greater than ever," *Middle East Institute*, September 9, 2022, https://www.mei.edu/publications/more-two-decades-911-threat-posed-jihadist-terrorism-greater-ever.

[41] Elizabeth Napolitano, "AI eliminated nearly 4,000 jobs in May, report says," *CBS News*, June 2, 2023, https://www.cbsnews.com/news/ai-job-losses-artificial-intelligence-challenger-report/.

[42] "1 in 3 Companies Will Replace Workers With AI in 2024," *Resume Builder*, November 8, 2023, https://www.resumebuilder.

suggest that up to 300 million jobs will be replaced by AI,[43] so this process is only likely to accelerate.

In others, the impacts may be environmental or contextual, such as generative AI's negative effects on an already fractured information environment. Freedom House's 2023 annual report found that generative AI tools had been used in at least 16 countries to distort information on social and political issues, and it concluded that moments of crisis or electoral periods can serve as flashpoints for AI-generated content.[44]

Simultaneously, a struggling media ecosystem is turning to generative AI to cut costs. Most newsrooms already use some form of AI in news production,[45] and some newly created news websites are hosting stories almost entirely written by AI software.[46] The impacts of these developments on news accuracy,[47] the huge imbalance in the resources devoted to detecting rather than producing AI-generated content, and the existing struggles faced by fact-checking organizations[48] have already made it difficult to determine what is real and what is fake online. As AI technology continues to develop, and the volume of inaccurate or biased AI-generated content increases, the information environment may deteriorate yet further. This has serious national and international security ramifications in its own right, but it will also benefit terrorists and violent extremist actors by promoting the type of conspiratorial, post-truth environment in which extremism of all kinds can flourish.

Finally, generative AI is likely to have second-order effects, ranging from bolstering authoritarian regimes to negatively impacting on climate change[49] (which can itself exacerbate the drivers of radicalization[50]) and perpetuating broader societal trends, including polarization, a rise in technology-facilitated gender-based violence,[51] and the embedding of biases and discrimination into decision-making.[52] Each of these

com/1-in-3-companies-will-replace-employees-with-ai-in-2024/.

[43] Chris Vallance, "AI could replace equivalent of 300 million jobs – report," *BBC*, March 28, 2023, https://www.bbc.com/news/technology-65102150.

[44] Allie Funk, Adrian Shahbaz, and Kian Vesteinsson, "Freedom on the Net 2023: The Repressive Power of Artificial Intelligence," *Freedom House*, October 3, 2023, https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence#generative-ai-supercharges-disinformation.

[45] Jade Drummond, "Newsrooms around the world are using AI to optimize work, despite concerns about bias and accuracy," *The Verge*, September 28, 2023, https://www.theverge.com/2023/9/28/23894651/ai-newsroom-journalism-study-automation-bias.

[46] Matthew Cantor, "Nearly 50 news websites are 'AI generated', a study says. Would I be able to tell?" *The Guardian*, May 8, 2023, https://www.theguardian.com/technology/2023/may/08/ai-generated-news-websites-study?utm_term=Autofeed&CMP=edit_2221&utm_medium=Social&utm_source=Twitter#Echobox=1683522052.

[47] Donie O'Sullivan & Allison Gordon, "How Microsoft is making a mess of the news after replacing staff with AI," *CNN*, November 2, 2023, https://edition.cnn.com/2023/11/02/tech/microsoft-ai-news/index.html?utm_source=substack&utm_medium=email.

[48] Tiffany Hsu & Stuart A. Thompson, "Fact Checkers Take Stock of Their Efforts. 'It's Not Getting Better,'" *The New York Times*, September 29, 2023, https://www.nytimes.com/2023/09/29/business/media/fact-checkers-misinformation.html.

[49] For example, an October 2023 study found that by 2027, AI server farms could be responsible for 0.5% of the world's energy demands. The data centers that power them already require astronomical amounts of water for cooling. For more information, see Victor Tangerman, "AI's Electricity Use Is Spiking So Fast It'll Soon Use As Much Power As an Entire Country," *The Byte*, November 10, 2023, https://futurism.com/the-byte/ai-electricity-use-spiking-power-entire-country.

[50] David Wells, "Climate Change, Terrorism, and Potential Implications for P/CVE," *Institute for Economics and Peace*, March 14, 2023, https://www.visionofhumanity.org/climate-change-terrorism-and-potential-implications-for-p-cve/.

[51] Rumman Chowdhury, "Technology-Facilitated Gender-Based Violence in an Era of Generative AI," *UNESCO*, November 2023, https://unesdoc.unesco.org/ark:/48223/pf0000387483/PDF/387483eng.pdf.multi.

[52] Garance Burke, Matt O'Brien, and *The Associated Press*, "Bombshell Stanford study finds ChatGPT and Google's Bard

trends is in itself arguably more significant than terrorism and violent extremism, but they will, in turn, also create conditions terrorists and violent extremists can exploit.

Conclusions and Recommendations

Given both the direct and indirect impacts of generative AI on terrorism and violent extremism, coordinating a coherent response will face significant challenges, particularly in the context of an AI arms race[53] and growing challenges to multilateralism and the rules-based order. Broadly speaking, these responses can be grouped into three areas: governmental, commercial, and academia or civil society.

National, Regional, and International Governmental Responses

There is a growing interest from governments and international organizations in understanding and responding to the risks and opportunities posed by AI, ranging from nascent legislation in the form of the European Union AI Act[54] to a variety of public-private sector consultations, bilateral and multilateral[55] agreements, and the creation of a UN High-Level Advisory Body on AI.[56]

Most of these initiatives are broad in scope, however, focusing on AI writ large rather than generative AI specifically, and few have focused to any great extent on terrorism, due to the greater salience of other AI-related threats and challenges along with the relative lack of data on generative AI and terrorism to date.

Yet there are some indications of growing interest in this specific issue. Europol published a Tech Watch Flash report in late March 2023, providing an overview of internal discussions on the exploitation of LLMs by criminals and malicious actors (including terrorists).[57] More recently, in September 2023, GIFCT concluded a red-team exercise focused on the impacts of generative AI on online terrorism and extremism.[58]

Further iterations of this type of multi-disciplinary exercise — particularly those that integrate the views of academia and civil society — will allow the counter-terrorism community to stay apprised of the evolving risks posed by new iterations of generative AI tools. National and regional bodies should also explore establishing dialogue between law enforcement agencies and generative AI companies, building on the frameworks and modalities established in law enforcement engagement with social media and other online platforms. Alongside these more collaborative, voluntary initiatives, governments may also need to consider specific counter-terrorism laws to address

answer medical questions with racist, debunked theories that harm black patients," *Fortune Well*, October 20, 2023, https://fortune.com/well/2023/10/20/chatgpt-google-bard-ai-chatbots-medical-racism-black-patients-health-care/.

[53] "Large language models: fast proliferation and budding international competition," *International Institute for Strategic Studies*, April 2023, https://www.iiss.org/publications/strategic-comments/2023/large-language-models-fast-proliferation-and-budding-international-competition/.

[54] Javier Espinoza, "EU agrees landmark rules on artificial intelligence," *Financial Times*, December 9, 2023, https://www.ft.com/content/d5bec462-d948-4437-aab1-e6505031a303.

[55] Raphael Satter & Diane Bartz, "US, Britain, other countries ink agreement to make AI 'secure by design,'" *Reuters*, November 28, 2023, https://www.reuters.com/technology/us-britain-other-countries-ink-agreement-make-ai-secure-by-design-2023-11-27/?utm_source=substack&utm_medium=email.

[56] "United Nations AI Advisory Body," *United Nations*, Accessed December 11, 2023, https://www.un.org/en/ai-advisory-body.

[57] "ChatGPT: The Impact of Large Language Models on Law Enforcement," *Europol*, March 27, 2023, https://www.europol.europa.eu/cms/sites/default/files/documents/Tech%20Watch%20Flash%20-%20The%20Impact%20of%20Large%20Language%20Models%20on%20Law%20Enforcement.pdf.

[58] GIFCT Red Team Working Group, "Considerations of the Impacts of Generative AI on Online Terrorism and Extremism," *GIFCT*, September 20, 2023, https://gifct.org/wp-content/uploads/2023/09/GIFCT-23WG-0823-GenerativeAI-1.1.pdf.

Photo above: US President Joe Biden walks to sign an executive order after delivering remarks on advancing the safe, secure, and trustworthy development and use of artificial intelligence, in the East Room of the White House in Washington, DC, on Oct. 30, 2023. Photo by BRENDAN SMIALOWSKI/AFP via Getty Images.

potential gaps in existing legislation presented by chatbots (and other forms of AI-generated content) in relation to criminal responsibility.

Tech Industry Responses

Past experience with social media platforms suggests that absent the meaningful threat of regulation or significant negative publicity regarding the misuse of its systems, cooperation and transparency from the generative AI sector is likely to be limited and piecemeal. Despite this prognosis, government engagement and broader advocacy efforts should continue, encouraging improved transparency of both the datasets used to train AI systems and the processes in place to prevent their misuse.

Simultaneously, industry and non-profit efforts to enhance the detection of AI-generated content should also be encouraged (and supported where possible). Efforts to integrate the generative AI sector into a variety of multilateral efforts to counter terrorist use of the internet (including GIFCT and the Extremism and Gaming Research Network[59]) should also continue. In this regard, OpenAI and Anthropic joined the Christchurch Call in November 2023,[60] a multi-sectoral initiative through which tech sector members make a series of non-binding commitments, including around transparency and preventing exploitation of their platforms by terrorists and violent extremists.

[59] "Extremism and Gaming Research Network," *EGRN*, Accessed December 11, 2023, https://extremismandgaming.org/.

[60] Ryan Heath, "Exclusive: New Zealand's Arden drafts AI in the fight against extremist content," *Axios*, November 13, 2023, https://www.axios.com/2023/11/13/ardern-ai-christchurch-call-openai-anthropic.

A Role for Academia and Civil Society

Finally, researchers and non-governmental groups also have a critical role to play. Further evidence-based research is needed on both the current and potential uses of generative AI by terrorists and violent extremists, in addition to potential responses. Civil society advocacy regarding the discrimination and bias inherent in many AI tools must continue, as should efforts to ensure that the necessary human rights protections are built into generative AI and the sector's relationships with government entities. Civil society organizations will also play a vital role in the creation and delivery of much-needed digital and media literacy education, helping to address both existing challenges and the new difficulty of differentiating between content that is AI generated versus created by humans.

Applying Lessons from Counter-Terrorism and P/CVE to Concerns Over Generative AI: Future Prospects

Across each of these sectors, stakeholders must remind themselves that while generative AI technology is new, many of the challenges it poses are not; moreover, many of the lessons learned over the past two decades of counter-terrorism and preventing and countering violent extremism (P/CVE) remain extremely relevant. These include the importance of multilateral cooperation, the centrality of both public-private partnerships and engagement with civil society organizations, and the need to respect human rights, particularly as AI and generative AI are likely to be fully integrated into counter-terrorism and P/CVE in the near future.

Assessing the prospects for these responses is difficult, particularly given the range of broader factors that will impact how big a problem they are addressing, including whether terrorists and violent extremists accelerate their adoption of the technology, and the extent to which the generative AI bubble is able to sustain its current rate of growth.

However, it seems unlikely that significant progress will be made in addressing the indirect impacts of generative AI on the conditions conducive to radicalization, given the breadth of the negative trends exacerbated by generative AI and the sheer volume of finance pushing this sector forward (making meaningful regulation difficult). This prognosis, although a negative one, further reinforces the need for holistic, whole-of-government and whole-of-society responses to terrorism and violent extremism in all its forms.

In contrast, the direct impacts of generative AI on the terrorist landscape remain limited at this stage (but are likely to accelerate), presenting a potential opportunity for multi-sectoral actors to understand and begin their response to an emerging issue before it grows out of control. There will undoubtedly be challenges, and the response will require a level of urgency and shared vision rarely seen in counter-terrorism policy circles; still, progress is achievable if meaningful action is taken soon.

Middle
East
Institute