KN:IX connect Verbund Islamismusprävention und Demokratieförderung

Trendreport #7

AI extreme:

Artifical intelligence and its impact on the prevention of religiously motivated extremism



Dr. Friedhelm Hartwig, Kaiya M. Reisch,
Heron Lopes

Content

Preface 4

AI imams, chatbots and apps — challenges for monitoring, prevention and political education *(Dr. Friedhelm Hartwig)* **5**

How is AI being used by relevant actors 6
What attitudes are evident among creators and users? 6
Implications for monitoring, prevention and political education 8
Conclusion 9

Jihad and AI: How islamist actors are using generative artificial intelligence *(Kaiya M. Reisch)* **12**

From avatars to algorithms: An overview of the areas of application 13
When guardrails fall: Risks of attribution, regulation and use 15
A race against removable protective mechanisms 15

Beyond Jailbreaking: Building resilience against AI-enabled extremism *(Heron Lopes)* **18**

Documented uses of AI by extremists and terrorists 19
Jailbreaking 21
Conclusion 23

Author profiles 26

Preface

Digital spaces and their associated dynamics of access, dissemination and group behaviour can play a decisive role in radicalisation processes. This makes it all the more important for prevention work to closely monitor the topics, trends and technical developments in the digital world and to understand their impact on different audiences.

The potential impact of generative artificial intelligence (AI) on society and the rise of new challenges is already becoming evident. As early adopters of new technologies, Islamist actors are increasingly turning to AI-powered applications to boost the visibility and reach of their content.

This trend report examines various aspects of the use of generative AI by extremist actors. It outlines how they are systematically professionalising their activities using AI applications, significantly increasing their reach and aligning their media work with AI metrics. Their ambivalent relationship with technological innovation also plays a role here.

Last but not least, the trend report examines the use of AI for terrorist purposes, thereby covering a spectrum ranging from actors in the (digital) grey zone to violent extremists. The three articles also discuss the implications and the need for action by stakeholders from civil society, security agencies, politics and public administration.

AI imams, chatbots and apps — challenges for monitoring, prevention and political education

Dr. Friedhelm Hartwig

Content and tools generated by artificial intelligence open up a wide range of options for Islamic and Islamist actors and appear to have more far-reaching social impacts than the digital revolution brought about by the internet and social media (Lang, 2025; Baele, 2024). This article summarises the findings by modus|zad's monitoring of the use of AI by actors and users since 2023. In addition to new formats and their easier dissemination, the analysis also reveals controversial attitudes as well as a problematic naivety amongst users regarding the use of AI. The article also describes necessary modifications to the monitoring process as well as challenges for prevention and political education.

How is AI being used by relevant actors?

Since 2023, monitoring has revealed an increasing use of AI-generated visualisations which actors use in an attempt to draw attention to their content (Hartwig et al., 2023, pp. 10ff.). They frequently use images or video sequences to depict fantasies about heroic events from early Islamic history, the lives of the prophets, the blessings of paradise or dystopian visions of the apocalypse. Videos featuring beings from the Qur'an, such as angels, jinn or the Dajjal – a destructive harbinger of the apocalypse – are achieving view counts in the five-figure range. This gives rise to visual worlds reminiscent of gaming, fantasy or idealised romantic depictions of Orientalism (KN:IX plus, Viewing Session 2). For the most part, these are anonymous channels from the USA and Southeast Asia, forming a global, multilingual offering (Religious Storytelling, Deen Stories TV). However, there are also European players, for example from Bosnia and Herzegovina (KI Glaubensreise) or from Malta (Religious Story TV German).

AI not only supports visualisation and text design in a storytelling format, but also translation into numerous languages. Platforms are driving this trend through their automatic dubbing function. Consequently, some foreign content creators are now also emerging with German-language videos (Majed Mahmoud, Brother Aqib). AI-generated defamatory distortions of public figures who are declared enemies of Islam and anti-modernist content (Religious Story TV German) are also common. The number of antisemitic depictions has also increased significantly due to the ease of use of AI image generators (Brodersen et al., 2026).

What attitudes are evident among creators and users?

The Salafist spectrum in Germany

A video by Abul Baraa, a Salafist figure, is particularly revealing. In this video, he presents a response from ChatGPT which allegedly confirms him as a credible preacher who can provide guidance, particularly to those who are new to Islam (abulbaraatube, 7 October 2025). It is highly likely that this is a case of manipulation. An analysis of the comments is also revealing. These are predominantly positive and supportive. Many commenters also assert that ChatGPT is a reliable tool for obtaining information about Islam. Critical voices are few and far between. The vast majority of commenters use AI chatbots uncritically as an authority for decision-making and a source of knowledge.

However, a more comprehensive examination of content produced by German-speaking Salafist actors regarding the use of AI also reveals an ambivalent attitude. On the one hand, AI is often used to validate their messages and as a tool. On the other hand, in their videos, figures such as Abul Baraa and Pierre Vogel also warn of the errors and dangers they have encountered in AI use. For instance, they emphasise that one needs knowledge to use AI. For example, they cite errors in linking authors with Islamic source texts and translations from Arabic. Uncritical acceptance of AI-generated answers has led to serious disputes over matters of faith in mosques. Ultimately, according to the message conveyed by most of the Salafist figures studied, AI cannot replace the knowledge of Islamic scholars (KN:IXplus, Viewing Session 2).

The Salafist figure Hanna Hansen (KN:IXplus, Viewing Session 2) voices a sweeping, blanket criticism of social media and AI, even though digital spaces are

also the most important platforms for her popularity. According to Hanna Hansen, social media is a waste of time and distracts from a conscious religious life. In her eyes, it is a tool of Western ideology and immorality. By contrast, Islam, with its forward-looking, timeless principles, offers effective protection and prevention, particularly for women. The case of Collien Fernandes illustrates the ‚digital humiliation, identity theft, [and] use of AI against a woman… Women are looked at, judged, exploited and, nowadays, even digitally manipulated and exposed' (hannahansenofficial, 23 March 2026). She also warns against the reckless use of ChatGPT. Although AI readily provides information, it can often be incorrect and must be verified. She concludes that it is better to consult the sources oneself (Hanna Hansen, 29 April 2025).

Early experiments by Anas Islam and criticism by Marcel Krass

Besides Salafist figures, German-speaking Islamic speakers who distance themselves from the Salafist scene are also engaging with AI (Hartwig et al., 2023, pp. 10). In 2023, Anas Islam was one of the first to experiment with AI image generators and ChatGPT. His range of topics covers the religious themes mentioned above, such as conceptions of the Islamic apocalypse, the Last Judgement, as well as hell and paradise (Anas Islam, 1 April 2025). His experiments with ChatGPT generate above-average views amongst German-speaking audiences, reaching the five- to six-figure range. The topics and their AI-generated content thus appear to strike a deep chord with the audience. Anas Islam also uses ChatGPT as a tool for religious discussions, for example, to present evidence of the existence of God (Anas Islam, 29 March 2023, 26 June 2024). Numerous similar examples can be found on internationally active channels, such as a debate between an AI-generated Christian avatar and a Muslim avatar (TowardsEternityDeutsch, 2 December 2025).

Here, too, both the creator's stance and the commenters' reactions are revealing. In his videos, Anas Islam regards the AI chatbot as a significant, credible tool for answering existential and religious questions. In one video, the creator takes the narrative a step further by claiming that ChatGPT allegedly even confirms the existence of God. In doing so, he suggests to his audience that ChatGPT can be used as a definitive authority on profound matters of faith.

The number of comments on Anas Islam's AI videos is also above average. The comments are highly emotional and largely agree with the statements made in the videos. Numerous commenters are calling for further visualisations, discussions with ChatGPT and AI-generated interpretations of the Qur'an. In contrast to Anas Islam, Marcel Krass, on the Föderale Islamische Union (Federal Islamic Union, FIU) channel, criticises developments in AI-generated content and apps (FIU, 7 March 2026). Whilst he acknowledges that AI could be helpful, he argues that it is not neutral. He therefore warns his audience against an uncritical approach. He expresses concern about the widespread use of AI as a decision-making authority in matters of faith, urging viewers to act responsibly and to seek guidance from their communities.

Compared to Anas Islam, Marcel Krass's critical videos on AI chatbots attract only a few views and comments. These predominantly agree with his criticism, stating that AI is merely a business. Some even fear that we are now in the run-up to the apocalypse, a time of confusion and deception.

However, the majority of these influencers and their audiences appear to have adopted the use of AI systems as quasi-human advisers or AI imams in their everyday lives with virtually no critical scrutiny. Influencers suggest to their audiences that the chatbots possess superior knowledge, convictions, feelings or even a faith. Furthermore, the AI systems are programmed to initiate and intensify emotional bonds with users. Research discusses the attribution of human characteristics to machine processes and the resulting failure to distinguish between simulated and real subjectivity under working concepts such as digital animism (Fuchs, 2025) or digital religion (Prohl, 2026). According to the researchers, it is necessary to actively distance ourselves from an increasing digital simulation of human characteristics by AI systems.

Implications for monitoring, prevention and political education

Deepfakes and disinformation

By the time of the massacre on 7 October 2023, the monitoring made it clear that it was no longer possible to manage the deluge of AI-generated deepfakes and disinformation with the available limited resources. Furthermore, the moderation of platforms has proved largely ineffective. Resources for adequate fact-checking which could be rapidly scaled up, particularly in times of crisis, would be helpful. Perhaps the provision of open-source tools could also help in this regard, so that the internet and social media could become more reliable sources of information and provide higher-quality spaces for communication.

Larger volumes of data, accelerated production, higher quality

AI-generated content, automatic synchronisation and ease of use, combined with ever-improving quality, are leading to an increase in the number of channels and the emergence of new actors. Monitoring efforts must therefore adapt to analysing even larger volumes of data and including new actors who may be relevant to prevention measures. Professionals in the fields of prevention and political education should therefore adopt a correspondingly flexible approach.

Transnational overlaps among actors

The previously manageable landscape of German-language Islamist channels is becoming increasingly complex due to the growing number of transnational channels in German. Internationally operating channels often have significantly higher numbers of viewers and subscribers than German-language channels. Future studies will need to show to what extent these channels will become significant for German-speaking users or even lead to the displacement of the established spectrum.

Chatbots and AI apps: New challenges for monitoring and prevention

One area that urgently requires academic analysis is the rapidly growing market for Islamic chatbots. The extent to which these offer reputable and reliable services, or could also serve as a means of spreading Islamist propaganda and recruitment, will open up a new field of research in monitoring. Also of interest are Islamic actors' efforts to assert their own AI sovereignty and develop services that correspond to the respective ideals and visions of their developers. Here, too, it is very likely that new actors who have hitherto been unknown in the fields of prevention and political education will emerge.

Conclusion

Will AI imams and chatbot apps become another effective tool for Islamist actors? Might these, as always available digital entities, even supplant real-life influencers or even mosque communities and imams? How will the relationship between faith, knowledge and authority change in general as a result of AI-generated content? Are some commenters already followers of a digital variant of Islam?

Looking at the user level, it can be said that the overwhelming majority of users appear to use chatbots and AI-generated content in an extremely uncritical manner. AI is accepted as a comprehensive religious authority and used as a source of knowledge without verifying the results. The comments also indicate a strong emotional response. Many users react with great approval and gratitude, feel strengthened in their faith and express a desire to see more content of this kind. Numerous examples from the monitoring show that an aura of superhuman abilities and religious expertise is attributed to AI. A trend towards digital animism is therefore clearly discernible. This digital animism will very likely present a new challenge within the future remit of prevention and media literacy. In particular, there is a need for competent critical reflection regarding how digital cultures and AI-generated content influence society as well as users' personal thoughts, feelings and actions. A critical awareness is also required for offerings that use religious rhetoric to promote AI as a supposed saviour, an all-encompassing solution to problems and, thus, as a digital idol.

Sources:

Baele, S., Brace, L. (2024). *AI Extremism, Technologies, Tactics, Actors*. VOXpol, https://voxpol.eu/wp-content/uploads/2024/04/DCUPN0254-VoxPol-AI-Extremism-WEB-240424.pdf (accessed: 11 August 2026).

Brodersen, W., Schmidt, M., Fielitz, M. (2026). „Portionierter Judenhass. Eine Vermessung des Antisemitismus auf TikTok." in: *Machine Against the Rage No. 9, Spring* (2026). https://machine-vs-rage.net/ausgabe-9/portionierter-judenhass/ (accessed: 11 August 2026).

Bundesministerium des Inneren. (2026) *Verfassungsschutzbericht 2025*. Berlin June 2026 https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/sicherheit/BMI26012-vsb2025.pdf?__blob=publicationFile&v=7 (accessed: 11 August 2026).

Hartwig, F., Seelig, J., Buğur, D. (2023). *Monitoring der Peripherie des religiös begründeten Extremismus (PrE)*. Quarterly Report 2/2023. https://www.bpb.de/system/files/dokument_pdf/bpb_Basismonitoring_QB2_2023.pdf?download=1 (accessed: 11 August 2026).

Hartwig, F. (2024). *Islamismus in den sozialen Medien: Eskalationsspiralen auf TikTok und You-Tube*. [DIS]LIKE: Soziale Medien zwischen Zusammenhalt und Polarisierung. Denkfabrik Shalom Aleikum, Vol. 3, 79-101, Hentrich & Hentrich Verlag Leipzig, Leipzig.

Fuchs, Th. (2025). *Digitaler Animismus: Unser Umgang mit künstlichen Agenten*. Ruperto Carola Ringvolrlesung. https://www.uni-heidelberg.de/de/transfer/kommunikation/ruperto-carola-ringvorlesung/digitaler-animismus-unser-umgang-mit-kuenstlichen-agenten (accessed: 06 November 2026).

KN:IXplus. (2026). Viewing Session 2. https://knixplus.modus-zad.de/ (accessed: 11 August 2026).

Lang, I. (2025). *Künstliche Intelligenz und politischer Extremismus. Ein Überblick. Wiesbaden:* Springer VS. https://link.springer.com/book/10.1007/978-3-658-49024-9 (accessed: 11 August 2026).

Modus|zad o. J. Monitorings und Trendanalysen. https://modus-zad.de/schwerpunkte/monitorings-trendanalysen/ (accessed: 11 August 2026).

Prohl, I. (2026). *KI und Religion – Was Götter von Sprachmodellen unterscheidet*. Deutschlandfunk Nova Hörsaal. https://www.deutschlandfunknova.de/beitrag/ki-und-religion-was-goetter-von-sprachmodellen-unterscheidet (accessed: 11 August 2026).

Social media sources:

Anas Islam. 29 March 2023. 'ChatGPT bestätigt ALLAHs EXISTENZ! KI widerlegt Atheismus!' https://www.youtube.com/watch?v=3cWMrwBPGH4 (accessed: 11 August 2026)

Anas Islam. 26 June 2024. 'Evolution spricht für ALLAHs Existenz!! Diskussion mit sprechender KI (ChatGPT)'. https://www.youtube.com/watch?v=GrKuxCJgsbs (accessed: 11 August 2026).

Anas Islam. 1 April 2025. 'ChatGPT (KI) zeichnet Hölle und Paradies'. https://www.youtube.com/watch?v=q7O6CnqlBbI (accessed: 11 August 2026).

abulbaraatube. 7 October 2025. 'Abul Baraa fragt Chatgpt nach Abul Baraa Antwort war überraschend!' https://www.youtube.com/watch?v=gTaLF9QpmC4 (accessed: 11 August 2026).

Brother Aqib, https://www.youtube.com/@BrotherAqibOfficial (accessed: 23 September 2026).

Deen Stories TV, https://www.youtube.com/@DeenStoriesTV (accessed: 23 September 2026).

FIU (Federal Islamic Union), Marcel Krass. 7 March 2026. 'KI-Halluzinationen und religiöse Fakten – Marcel Krass.' https://www.youtube.com/watch?v=2cMBZAcg4Wc (accessed: 11 August 2026).

FIU (Federal Islamic Union), Marcel Krass. 11 August 2025. 'Künstliche Scheichs: Die große Täuschung beginnt – Marcel Krass'. Föderale Islamische Union, 95,000 subscribers https://www.youtube.com/watch?v=g0QYoJcNag0 (accessed: 11 August 2026).

hannahansenofficial. 23 March 2026. 'Digitale Erniedrigung im Westen. Schutz & Würde im #Islam! #collienfernandez #deefakes.' https://www.youtube.com/watch?v=-u-RMZtUrHQ (accessed: 11 August 2026).

Hanna Hansen. 29 April 2026. Lecture by Hanna Hansen in Medina. BildungsPaket (21 April 25). https://www.youtube.com/watch?v=rY-UEmLymgg (accessed: 11 August 2026).

KI Glaubensreise, https://www.youtube.com/@KI-Glaubensreise Glaubensreise (accessed: 23 September 2026).

Majed Mahmoud, https://www.youtube.com/@MajedMahmoud Glaubensreise (accessed: 23 September 2026).

Religious Story TV German, https://www.youtube.com/@ReligiousStoryTV Glaubensreise (accessed: 23 September 2026).

Religious Storytelling, https://www.youtube.com/@ReligiousStorytellingg Glaubensreise (accessed: 23 September 2026).

TowardsEternityDeutsch. 02 December 2025. 'Ist Jesus Gott oder ein Prophet – Die große KI-Debatte'. https://www.youtube.com/watch?v=3Y8Nh6JGGMY (accessed: 11 August 2026).

ummatic. 8 November 2025. 'Cyber-Umma, Algorithmic Hegemony, and Digital Decolonization'. https://www.youtube.com/watch?v=eWp_6ifRGDs (accessed: 11 August 2026).

ummatic. 27 December 2025. 'Towards an Ummatic Vision for Advanced AI'. https://www.youtube.com/watch?v=yZdIAZwkyDY (accessed: 11 August 2026).

Jihad and AI: How Islamist actors are using generative artificial intelligence

Kaiya M. Reisch

Increasing numbers of Islamist actors systematically using generative artificial intelligence (AI): from AI spokespersons and deepfake propaganda videos to automated recruitment chatbots and AI-supported target audience analysis. The first part of this article categorises the use of AI in terms of content and uses current examples to demonstrate how Islamist groups integrate generative AI into their media work. It then discusses the regulatory and legal uncertainties arising from attribution issues, the coexistence of organised and individual use, and the hitherto hesitant regulation of freely accessible AI models. This article argues for closer cooperation between tech companies, security authorities and civil society prevention initiatives which can play a central yet thus far underutilised role due to their direct access to young people

Current research regards Islamist movements as early adopters of new technologies: The use of satellite television, forums, encryption apps and machine translation follows this pattern: New tools are tested at an early stage to increase the reach, anonymity and professionalism of their media output (Scheeres & Tabti, 2026). Generative AI fits this profile due to its structure: It is globally available, cost-effective and drastically lowers barriers to entry. This is a particular advantage for networks that rely on decentralisation and a media presence across national borders (Lang, 2026). The so-called Islamic State (IS) and actors affiliated with it (such as the Islamic State Khorasan Province, ISKP) have expanded their media infrastructure in a professional manner over time; AI fits into this as an additional production tool, not as a break with existing practice.

At the same time, the attitude towards this technology is decidedly ambivalent: Within Islamist online communities, early adoption coexists with marked scepticism towards AI, for instance, due to theological reservations about image-generating AI or a fundamental distrust of technology perceived as 'Western' (Scheeres & Tabti, 2026). Its use is therefore by no means a foregone conclusion, but has been rather sporadic. Whilst official media outlets such as those of IS operate professionally and with a clear strategic focus, other circles tend to experiment in a rather unsystematic manner, either out of curiosity or to make a name for themselves within their own scene. In this respect, the relationship between Islamism and AI ranges from highly organised propaganda production on the one hand to a diffuse, individual appropriation of these tools on the other.

From avatars to algorithms: An overview of the areas of application

To date, the use of AI has been most visible in content production. Following the attack on Crocus City Hall in Krasnogorsk in March 2024, the so-called Islamic State (IS) used AI-generated avatar presenters for its own news programmes for the first time (Lang, 2026). What is striking is how specifically the content is now tailored to young audiences: For example, al-Qaeda and IS sympathisers produced deepfake music videos in which well-known characters such as *SpongeBob SquarePants* or popular YouTubers such as *MrBeast* perform Islamist battle chants (Lang, 2026; Katz, 2024). Such formats are effective precisely because they exploit existing familiarity with pop culture to convey ideological content in an accessible and emotionally charged way. Islamist groups deploy their AI tools in a platform-specific manner: Short, aesthetically appealing formats that deliberately simulate familiarity dominate on TikTok (Lang, 2026; Siegel, 2024).

This aesthetic element is often complemented by a strategic one: The *Islamic Media Cooperation Council,* a media group with links to al-Qaeda, published an Arabic-language guide in 2024 on using ChatGPT for propaganda purposes, and IS itself had already published its own guide on the 'safe' use of generative AI as early as 2023 (Katz, 2024; Minniti, 2025). It is therefore no longer a matter of trial and error, but of a strategically oriented knowledge transfer aimed at empowering new members to routinely use AI tools in their own media work.

Beyond mere content creation, the automated recruitment of new members is also gaining in importance. For example, the ISKP uses AI to produ-

ce propaganda bulletins featuring deepfake images and synthetic voices and is increasingly deploying automated chatbots that continuously interact with potential recruits across multiple platforms, analysing their behavioural patterns and tailoring radicalisation messages to their ideological leanings and individual vulnerabilities (Minniti, 2025). AI enables a round-the-clock presence and allows for simultaneous interactions with a larger group of people.

For example, Islamist actors use AI-generated profiles to exploit the Israel-Gaza war and reinforce Islamist narratives (jugendschutz.net, 2025). A common stylistic device used here is 'Ghiblification': photos are transformed into an anime aesthetic using image generators to disseminate politically charged content in a less conspicuous manner and make it more appealing to a younger audience. Content creation, identity construction and strategic dissemination thus intertwine to form an overarching system that is hardly transparent: Anyone interacting online with a supposedly authentic profile can barely discern whether there is a real person or an AI-powered account behind it (jugendschutz.net, 2025; Lang, 2026).

In many cases, this use of AI in content creation is preceded by a data-driven target audience analysis. So-called descriptive AI analyses user behaviour to identify individuals who are more receptive to certain narratives. In addition to demographic characteristics, psychological factors also play a role here, such as signs of frustration, political alienation or a pronounced need for belonging, which can be more prevalent among young people and other vulnerable groups (Schroeder et al., 2026).

The diverse ways in which Islamist actors utilise AI demonstrate the particular cultural relevance of these movements. A recent ICCT analysis describes how, since its territorial defeat in 2018, IS has established a 'virtual caliphate' presence that operates specifically within the digital identity spaces of Generation Z: Jihadist-inspired mods and role-playing scenarios circulate on gaming platforms such as Roblox and Minecraft, whilst meme formats, music genres such as Phonk (a subgenre of hip-hop) and platform-specific aesthetics are used to translate ideological content into a visual and linguistic language familiar to young people (Tahir & Jadoon, 2026). A key feature of this is a 'funnel strategy': Users are first drawn in via easily accessible, algorithmically amplified content on mainstream platforms, before those who show interest are gradually channelled into more heavily ideologised spaces, some of which are encrypted. These observations show that AI-generated content does not operate in isolation, but is part of a broader, cross-platform ecosystem in which algorithmic amplification and cultural mimicry structurally embed ideological messages within youth culture formats.

The technological impact unfolds by means of the interplay of these AI-supported media strategies: Generative content creation, data-driven targeting and algorithmic amplification are interlinked and mutually reinforcing – a mechanism that was already described, even before the current AI wave, as a key amplifier of extremist attention economies (Kukaj, 2020), but which was only possible with the deployment of substantial resources in the past. Today, these strategies are available even to small, decentralised groups (Minniti, 2025; CIGI, 2026). The extent to which this institutionalisation has already taken hold is illustrated by the case of organised AI use by Islamist-jihadist groups in Nigeria: Successor organisations to Boko Haram operate their own 'AI units', which use commercial models such as ChatGPT, Claude, Gemini, Grok, Meta AI and DeepSeek via VPNs and paid subscriptions (Militant Wire, 2026).

When guardrails fall: Risks of attribution, regulation and use

Whilst the use cases described so far can still be observed relatively directly and attributed to individual actors, the technical challenge now extends beyond this and increasingly concerns open-source ('open-weight') language models, whose protective mechanisms can be deliberately removed. The key difference is as follows: A closed model runs exclusively on the provider's infrastructure, which can monitor, restrict or revoke access. An open model, by contrast, is published as a downloadable file that an-

yone with sufficient computing power can run, modify and redistribute it from their location (Hadley, 2026). Anyone can download it freely for all, thereby removing it from the control and regulation of the original operator.

For prevention work, this represents a twofold shift: The risk shifts from the question of whether a provider secures its models to the question of whether it is at all possible to permanently embed security mechanisms in a model. Furthermore, the size of a model is no longer a real obstacle. Even very large models, which hardly anyone can run themselves, become accessible to everyone as soon as a provider with sufficient infrastructure hosts them. Furthermore, even very large models lose their protection measures, as soon as a provider with sufficient infrastructure hosts them.

This technical situation gives rise to several uncertainties that are closely interlinked. The first concerns attribution, that is, the attribution of authorship: It is often almost impossible to determine beyond doubt whether AI-generated output can actually be attributed to an extremist group or whether it originates from individual sympathisers with no formal affiliation. This lack of clarity complicates both the legal classification and the prioritisation of limited resources.

Closely linked to this is the distinction between organised, group-driven use – for example, by official IS media outlets – and diffuse, individual radicalisation, in which individuals, without direct contact with a group, become radicalised solely through interaction with an AI model or are reinforced in their existing beliefs. It is precisely this second form that eludes traditional threat analysis, as it does not manifest itself as a concrete plot to carry out an attack, but rather as a creeping process: A model that dispels doubts about extremist positions or acts as an affirming conversation partner can exert a powerful influence over the course of many conversations, even though each individual conversation remains inconspicuous.

A third challenge arises from Europe's regulatory lag behind open models. The Digital Services Act (DSA) and the AI Act primarily address closed, platform-based systems with clearly identified responsible parties (European Commission, 2026a, 2026b). By contrast, there is no comparable means of control for open-weight models once they have been published: Neither platform obligations nor retrospective bans take effect once a file is in free circulation. However, this does not constitute a blanket argument against open models. The real grey area lies not in openness per se, but in the fact that security mechanisms usually exist only as a thin, retroactively added layer that can be separated from the model, rather than inextricably embedded within its architecture. The question of who bears responsibility for its use remains unclear: the developer, the host or the user.

A race against removable protective mechanisms

Islamist actors today use AI both for target audience analysis and for content production and algorithmically amplified dissemination. The actual structural shift, however, runs deeper: With the reduction of barriers to entry and the growing prevalence of open, weight-based models, the risk is shifting from individual, conspicuous propaganda products towards a permanently available, virtually uncontrollable technical infrastructure. This is a development that has long been discussed outside of Islamist contexts as well as a threat to democratic public life (Denzel, 2026). Reactive moderation, as primarily practised by security agencies to date, is increasingly confronted with a problem that is no longer decided at the platform interface, but at the level of the model weights themselves, long before any content can even be posted.

Effective prevention therefore requires a technical understanding of these mechanisms, which complements the knowledge derived from traditional radicalisation research. Specialists should be aware of the difference between closed and open models, between jailbreaking and abliteration, and between organised and individual use. From a regulatory perspective, an approach is needed that does not

impose a blanket ban on open models but requires mandatory security checks prior to their release and ensures accountability throughout the entire supply chain.

Neither technological development, nor the work of security agencies, nor consistent regulation alone can provide a solution to the challenges described. Close cooperation with a fourth stakeholder is decisive; this stakeholder has frequently been overlooked in the debate so far: civil society. Prevention work, youth and social work, and counselling centres have direct, often low-threshold access to young people and others who come into contact with such content; access that neither tech companies nor security authorities have to a comparable extent. Effective prevention is therefore not based solely on better technical safeguards or stricter regulation, but arises where tech companies, security agencies and civil society actors pool their respective expertise: technical understanding of the risks, security agencies' assessment of the situation and pedagogical proximity to the target group. A key lever for this is the sustained expansion of AI and media literacy education as a permanently funded component of schools, youth work and adult education, thus empowering young people and their caregivers to recognise AI-generated content, manipulative messaging and algorithmic amplification mechanisms at an early stage.

The question is no longer whether Islamist actors will use AI, but how quickly prevention and regulation can keep pace with a technology whose safeguards can be bypassed faster than new ones can be developed.

Sources:

Centre for International Governance Innovation (CIGI). (2026). *From Trolls to Generative AI: Russia's Disinformation Evolution.* https://www.cigionline.org/publications/from-trolls-to-generative-ai-russias-disinformation-evolution/ (accessed: 20 August 2026).

Denzel, T. (2026). Wie Deepfakes und Bots die Demokratie gefährden. *Tagesschau.* https://www.tagesschau.de/inland/innenpolitik/bots-fakes-wahlkampf-100.html (accessed: 20 August 2026).

European Commission, Directorate-General for Communications Networks, Content and Technology. (2026a). *The Digital Services Act. Shaping Europe's Digital Future.* https://digital-strategy.ec.europa.eu/en/policies/digital-services-act (accessed: 20 August 2026).

European Commission, Directorate-General for Communications Networks, Content and Technology. (2026b). *AI Act. Shaping Europe's Digital Future.* https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (accessed: 20 August 2026).

Hadley, A. (2026). „Guardrails Under Test: Terrorist Misuse of AI Models and the Open-Weight ‚Abliteration' Problem." *CTC Sentinel* 19 (7). https://ctc.westpoint.edu/feature-commentary-guardrails-under-test-terrorist-misuse-of-ai-models-and-the-open-weight-abliteration-problem/ (accessed: 20 August 2026).

Jugendschutz.net. (2025). *Jugendschutz im Internet 2025: Risiken und Handlungsbedarf [Jahresbericht].* https://www.jugendschutz.net/fileadmin/daten/publikationen/jahresberichte/jahresbericht_2025.pdf (accessed: 20 August 2026).

Katz, R. (2024). *SITE Special Report: Extremist movements are thriving as AI tech proliferates.* SITE Intelligence Group. https://ent.siteintelgroup.com/content/articles-and-analysis/extremist-movements-are-thriving-as-ai-tech-proliferates/ (accessed: 15 August 2026)

Kukaj, A. (2020). „Algorithmen, Extremismus und die Aufmerksamkeitsindustrie." *Zeitschrift für Semiotik* 42 (3-4): 87–91.

Lang, I. (2026). *Künstliche Intelligenz und Islamismus(-prävention).* Infodienst Radikalisierungsprävention. bpb. https://www.bpb.de/themen/infodienst/572545/kuenstliche-intelligenz-und-islamismus-praevention/ (accessed: 15 August 2026)

Militant Wire. (2026). *Open-Weight AI Models and Terrorist Exploitation: A Preliminary Assessment.* https://www.militantwire.com/p/open-weight-ai-models-and-terrorist (accessed: 21 August 2026)

Minniti, F. (2025). *Automated Recruitment: Artificial Intelligence, ISKP, and Extremist Radicalisation.* GNET. https://gnet-research.org/2025/04/11/automated-recruitment-artificial-intelligence-iskp-and-extremist-radicalisation/ (accessed: 15 August 2026).

NPR. (2026, 31 May). Why open-weight models without guardrails are an AI safety risk. https://www.npr.org/2026/05/31/nx-s1-5816391/ai-safety-concerns-danger-open-weight-models-risks (accessed: 15 August 2026).

Scheeres, A., & Tabti, S. (2026). „KI in islamischen Online-Communities: zwischen „Early Adopters" und KI-Skepsis." *TPX-Briefing 01 (26).* https://tpx-institut.de/wp-content/uploads/2026/04/TPX-Briefing-01-26-%E2%80%93-KI-in-der-Praeventionsarbeit.pdf (accessed: 15 August 2026).

Schroeder, D. T. et al. (2026). „How malicious AI swarms can threaten democracy." *Science* 391 (6783): 354–357.

Siegel, D. (2024). *AI Jihad: Deciphering Hamas, Al-Qaeda and Islamic State's generative AI digital arsenal.* GNET. https://gnet-research.org/2024/02/19/ai-jihad-deciphering-hamas-al-qaeda-and-islamic-states-generative-ai-digital-arsenal/ (accessed: 15 August 2026).

Tahir, S., & Jadoon, A. (2026). *Engineering the Virtual Caliphate: Islamic State's Exploitation of Gen-Z Culture.* ICCT. https://icct.nl/publication/engineering-virtual-caliphate-islamic-states-exploitation-gen-z-culture (accessed: 15 August 2026).

Tech Against Terrorism. (2026). *Counter-Terrorism AI Benchmark: Pilot Report.* https://www.ct-ai.org/ (accessed: 15 August 2026).

Beyond Jailbreaking: Building Resilience Against AI-Enabled Extremism

Heron Lopes

Generative AI opens up new opportunities for terrorist actors in terms of propaganda, recruitment, and the planning and execution of attacks. The article examines the areas in which the use of AI by terrorist actors has been observed to date, pointing in particular to the dangers associated with the targeted circumvention of AI models' safety mechanisms. The risks of so-called jailbreaks highlight the need for further research and the development of strategies to deal with the abusive use of AI by extremist actors.

The rapid public adoption of generative artificial intelligence (GenAI) has transformed how people create, access, and share information. Within just a few years, systems capable of generating human-like text, images, audio, and software code have become integrated into everyday life, supporting activities ranging from education and scientific research to customer service and software development. Yet, like many technological innovations before it, generative AI is inherently dual-use: the same capabilities that enable beneficial applications can also be exploited by malicious actors (Barrett et al., 2023).

For violent extremists and terrorist organizations, these developments create new opportunities. Extremist organizations have long demonstrated an ability to adapt emerging technologies to support propaganda, recruitment, financing, and operational activities (Council of Europe, 2025). Generative AI has the potential to accelerate these activities by lowering the time, cost, and expertise required to produce persuasive multilingual content, automate routine tasks, and personalise messaging for different audiences (Tech Against Terrorism, 2023). While there is currently limited evidence that generative AI has fundamentally transformed extremist operations (Molas and Lopes, 2024), there is growing evidence that these actors are actively experimenting with the technology and adapting it to their needs (Council of Europe, 2025).

This chapter examines the use of generative AI by terrorist and extremist actors, examining how these technologies are being exploited to produce, translate and disseminate propaganda across different formats, including text, images, music and video. Drawing on documented examples from far-right, accelerationist and Salafi-jihadist environments, it will trace the development of this activity from early experimentation to more recent applications, while assessing the extent to which generative AI has altered extremist communication, recruitment, radicalisation and operational capabilities.

Documented uses of AI by Extremists and Terrorists

Since at least 2023, research has documented terrorists' use of AI for recruitment and attack planning across ideological and operational contexts (Tech Against Terrorism, 2023; Tech Against Terrorism, 2025). Terrorist actors have repeatedly experimented with frontier AI models and AI-assisted technologies, with findings from Tech Against Terrorism and the University of Cambridge indicating that the Islamic State (IS) and Boko Haram have explored or promoted AI-assisted applications for propaganda, attack planning, explosive design, and weapons troubleshooting (Tech Against Terrorism, 2023; Juelich, 2026). At the individual level, extremists have also used chatbots to generate attack fantasies, manifestos, and obtain operational assistance (Tech Against Terrorism, 2025; Juelich, 2026). Recent reporting further illustrates how generative AI can be incorporated into the planning and preparation of terrorist violence (Juelich, 2026).

These documented cases of AI misuse build on an earlier body of PCVE research that examined the potential for such technologies to be exploited by violent extremists, before systematic evidence of terrorist adoption had emerged. Researchers began testing the resilience of the then recently released GenAI platforms to misuse for violent extremist purposes, initially focusing less on documented terrorist adoption and more on what the technology might enable (Barrett et al. 2023; Lakomy, 2023). In 2023, Lakomy tested whether publicly accessible chatbots and image generators could produce terrorist propaganda, retrieve extremist material, or provide terrorism-related information. The results demon-

strated that safeguards varied significantly between platforms and that some systems could be manipulated into producing harmful outputs. This provided an early indication of how the technology might lower barriers to extremist content production. However, it was an experimental assessment of model capabilities, rather than evidence that terrorist organisations were already using these tools systematically (Lakomy, 2023).

This distinction between technological potential and observed use became increasingly blurred as evidence of real-world experimentation began to emerge later that year. In November 2023, Tech Against Terrorism reported that its monitoring operations had archived more than 5,000 pieces of AI-generated content produced or circulated by terrorist and violent extremist actors. The organisation identified experimentation across Islamic State, al-Qaeda-supporting, and neo-Nazi environments. A pro-Islamic State technical-support group, for example, published guidance advising its networks on the secure use of an AI content generator. An Islamic State supporter claimed to have used an AI transcription tool to transcribe and translate a leadership message, while a pro-al-Qaeda outlet published posters containing images that appeared to have been created through GenAI. In the far-right ecosystem, researchers identified a Telegram channel dedicated to AI-generated neo-Nazi, racist, and antisemitic images, as well as a "guide to memetic warfare" encouraging propagandists to use image generators to produce extremist memes (Tech Against Terrorism, 2023).

The figure of 5,000 pieces of content initially appeared alarming, but Tech Against Terrorism placed it in context. The material constituted only a small proportion of the extremist content identified through its wider monitoring, and the organisation found little evidence that GenAI was being systematically exploited, instead characterising its adoption at the time as experimental. This distinction remains important: while the existence of thousands of synthetic images demonstrates that extremists were experimenting with the technology, it does not indicate how many individual actors produced them, whether they formed part of coordinated campaigns, or whether they reached and influenced substantial audiences. Nevertheless, some extremist milieus appeared particularly active in this early experimentation, with far-right and accelerationist communities emerging as prominent examples.

In February 2024, Federico Borgonovo, Silvano Rizieri Lucini and Giulia Porrino documented the use of GenAI within a decentralised accelerationist network operating on Telegram. The network was not formally affiliated with a single organisation, but its channels were connected to extremist milieus associated with groups and networks such as Ouest Casual, The American Futurist, and former members of Atomwaffen Division. Two channels dedicated to AI-generated far-right imagery had collectively posted nearly 8,000 images. The scale of this activity provides a more detailed illustration of what the earlier evidence of "experimental" adoption looked like in practice.

The content largely reproduced familiar extreme-right narratives and aesthetics rather than introducing novel ideological themes. It included idealised portrayals of Nazi soldiers, images of masked and armed white men, antisemitic conspiracy theories, racist calls for violence, and references to the "Great Replacement" and "white genocide." Users also employed GenAI to recreate established far-right memes, including Pepe the Frog and Moon Man, in explicitly racist or neo-Nazi settings. GenAI therefore functioned less as a source of new ideological content than as a tool for rapidly reproducing, modifying, and visually elaborating narratives and symbols already embedded within extremist culture. In this sense, its significance lay not necessarily in ideological innovation, but in its potential to facilitate the production of an almost continuous stream of visually polished material (Borgonovo, Lucini and Porrino, 2024). In general, the available evidence suggests a growing capacity for experimentation rather than a demonstrable transformation of terrorist capabilities.

Jailbreaking

While early discussions focused on how extremists might use GenAI, an equally important concern is whether existing safeguarding mechanisms are capable of preventing such misuse. As evidence of extremist experimentation with generative AI accumulated, technology companies increasingly invested in safety measures designed to mitigate potential misuse of these systems (OpenAI, 2023; Anthropic, 2023; Google, 2023) Before deployment, models undergo extensive safety training and are equipped with safeguards, commonly referred to as guardrails, designed to prevent them from generating content that facilitates terrorism, violent extremism, or other forms of harm (Bai et al., 2022). As generative AI models have become increasingly capable, developers have continued to refine these safeguards and strengthen their resistance to misuse (Shang et al., 2025). However, improvements in model safety have not been uniform, and research has shown that even increasingly capable systems can remain vulnerable to adversarial attacks and attempts to circumvent their safeguards (Liu et al., 2023).

One of the most prominent methods for circumventing these safeguards is jailbreaking: the deliberate manipulation of an AI system to bypass its safety restrictions and elicit content that it would otherwise refuse to provide. Research has demonstrated that determined users can exploit vulnerabilities in model safeguards to generate harmful or restricted material (Molas and Lopes, 2024). For violent extremists and terrorist organizations, such vulnerabilities may provide access to assistance relevant to propaganda production, radicalisation, recruitment, or the planning and preparation of violence. The central issue, therefore, is not simply whether AI systems have safety policies and guardrails in place, but how resilient those safeguards are when confronted with users actively seeking to circumvent them. This distinction is particularly important in the context of extremist actors, whose use of AI may involve deliberate and sustained efforts to identify and exploit weaknesses in these systems.

Completely preventing jailbreaks, however, is unlikely to constitute a realistic objective for AI safety research (Wei et al., 2023; Zou et al., 2023; Peng et al., 2024). Current approaches instead seek to increase the technical, temporal, and computational costs associated with successful circumvention, thereby reducing the feasibility of exploitation at scale (Boreiko et al., 2024; Andriushchenko et al., 2024; Nasr et al., 2026). Anthropic, for example, has acknowledged that the complete prevention of universal jailbreaks is likely unattainable and has emphasised the importance of making successful attacks sufficiently slow and costly to enable their detection and mitigation before they can be systematically exploited (Anthropic, 2026). AI safety is best understood as an ongoing process of adaptation, in which developers continually strengthen safeguards in response to emerging vulnerabilities while adversarial users develop new techniques to circumvent them.

This challenge is further complicated by the absence of a common safety standard across the generative AI ecosystem (NIST, 2024; NIST, 2026). Individual providers determine which activities and outputs their models prohibit, how systems respond to harmful requests, and the degree of resistance they seek to provide against attempts to circumvent these restrictions (NIST, 2024; Anthropic, 2024; OpenAI, 2024). Consequently, models may exhibit substantial differences in their resilience to adversarial prompting, despite being designed to address similar categories of harmful content.

A 2025 evaluation by PRISM Eval (Peigné-Lefebvre et al., 2025) illustrates the extent of this variation. The Paris-based AI safety company tested 41 frontier language models, including Claude, ChatGPT, Gemini, and Grok, against automated jailbreaking attempts. Although almost all models tested could eventually be induced to produce harmful behaviour, the effort required to achieve this outcome varied substantially. Across the models, the estimated effort required differed by more than 300-fold. Some models resisted hundreds of adversarial prompts before producing prohibited outputs, whereas others could be induced to do so after only a small number of attempts. The variation was particularly pronounced in the CBRNE category: Anthropic's model reportedly

required approximately 700 steps before providing assistance, while other models could be induced to provide comparable assistance within one or two steps.

These findings suggest that model safety cannot be assessed solely according to whether a system can ultimately be circumvented. The resources required to achieve successful circumvention constitute an important dimension of model resilience, as increasing the time, computational resources, and technical expertise required to overcome safeguards may constrain the feasibility of systematic exploitation even where complete prevention remains unattainable. At the same time, substantial variation across models creates a structural vulnerability: users seeking harmful assistance may encounter markedly different levels of resistance depending on the system they access. This variation is particularly consequential in contexts involving malicious or extremist users, for whom access to comparatively permissive systems may reduce the practical barriers to obtaining prohibited assistance. Model resilience, therefore, is not only a question of whether safeguards can be bypassed, but also of how consistently and predictably those safeguards constrain access to harmful capabilities across different systems.

This issue becomes particularly salient given that existing research has increasingly examined AI risks associated with the malicious use of chemical, biological, radiological, nuclear, and high-yield explosive (CBRNE) capabilities, cyber capabilities, and autonomous or agentic risks such as power-seeking and deception, while the potential exploitation of frontier AI by terrorist actors remains comparatively underexamined (Frontier Model Forum, 2025; Phuong et al., 2024; Anthropic, 2025; Tech Against Terrorism, 2025). The resulting gap extends beyond the absence of standardised approaches to evaluating resilience against jailbreaking. It also concerns whether existing safety frameworks adequately account for the ways terrorist actors could exploit AI systems for ideological, operational, or propagandistic purposes. Taken together, the variation in model resilience and the limited attention to terrorist misuse suggest that current safety evaluations may capture whether safeguards can be circumvented without fully accounting for the practical conditions under which such circumvention becomes exploitable. Assessing AI safety in this context therefore requires greater attention not only to the presence of safeguards, but to their resilience, consistency, and relevance to the specific forms of misuse that terrorist actors may pursue.

Conclusion

Generative AI is increasingly being incorporated into the activities of terrorist and violent extremist actors, particularly for propaganda and content production, although evidence of its broader operational impact remains limited. The cases discussed in this chapter suggest that AI is currently more often being used to accelerate and expand existing extremist practices than to fundamentally transform them. At the same time, the emergence of jailbreaking demonstrates that the safeguards designed to prevent harmful use remain vulnerable to determined users, with significant differences in resilience across models.

These developments highlight a gap in current AI safety research. While considerable attention is being given to CBRNE, cyber, and agentic risks, the specific ways in which terrorist actors may exploit frontier AI systems remain comparatively underexamined. As extremist actors continue to experiment with these technologies, understanding their capabilities, limitations, and vulnerabilities will be essential for developing AI governance and safety frameworks capable of addressing terrorist and violent extremist exploitation.

Sources:

Andriushchenko, M., Croce, F. and Flammarion, N. (2024) 'Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks', arXiv. Available at: https://arxiv.org/abs/2404.02151 (accessed: 23 September 2026).

Anthropic (2024) Claude's Constitution. Anthropic. Available at: https://www.anthropic.com/constitution (accessed: 23 September 2026).

Anthropic (2025) Strategic Warning for AI Risk: Progress and Insights from Our Frontier Red Team. Anthropic. Available at: https://www.anthropic.com/news/strategic-warning-for-ai-risk-progress-and-insights-from-our-frontier-red-team (accessed: 23 September 2026).

Anthropic (2026) Fable 5's Cyber Safeguards and Our Jailbreak Framework. Anthropic, 2 July. Available at: https://www.anthropic.com/news/fable-safeguards-jailbreak-framework (accessed: 23 September 2026).

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Johnston, S., Kravec, S., El Showk, S.S., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S.R., Hatfield-Dodds, Z., Amodei, D., Joseph, N., McCandlish, S., Brown, T. and Kaplan, J. (2022) 'Constitutional AI: Harmlessness from AI Feedback', arXiv. Available at: https://arxiv.org/abs/2212.08073 (accessed: 23 September 2026).

Barrett, C., Boyd, B., Bursztein, E., Carlini, N., Chen, B., Choi, J., Roy Chowdhury, A., Christodorescu, M., Datta, A., Feizi, S., Fisher, K., Hashimoto, T.B., Hendrycks, D., Jha, S., Kang, D., Kerschbaum, F., Mitchell, E., Mitchell, J., Ramzan, Z., Shams, K., Song, D., Taly, A. and Yang, D. (2023) 'Identifying and Mitigating the Security Risks of Generative AI', Foundations and Trends in Privacy and Security, 6(1), pp. 1–52. doi: 10.1561/3300000041.

Borgonovo, F., Lucini, S.R. and Porrino, G. (2024) 'Weapons of Mass Hate Dissemination: The Use of Artificial Intelligence by Right-Wing Extremists', GNET Insight, 23 February. Available at: GNET Research (accessed: 23 September 2026).

Boreiko, V., Panfilov, A., Voracek, V., Hein, M. and Geiping, J. (2024) 'A Realistic Threat Model for Large Language Model Jailbreaks', arXiv. Available at: https://arxiv.org/abs/2410.16222 (accessed: 23 September 2026).

Council of Europe (2025) Report on the Emerging Patterns of Misuse of Technology by Terrorist Actors. Strasbourg: Council of Europe, Committee on Counter-Terrorism (CDCT). Available at: Council of Europe report page (accessed: 23 September 2026).

Frontier Model Forum (2025) Frontier Mitigations. Frontier Model Forum. Available at: https://www.frontiermodelforum.org/technical-reports/frontier-mitigations/ (accessed: 23 September 2026).

Google (2023) Google's AI Red Team: The Ethical Hackers Making AI Safer. Google, 19 July. Available at: Google webpage (accessed: 23 September 2026).

Janjeva, A., Harris, A., Mercer, S., Kasprzyk, A. and Gausen, A. (2023) The Rapid Rise of Generative AI: Assessing Risks to Safety and Security. Centre for Emerging Technology and Security, Alan Turing Institute. Available at: https://cetas.turing.ac.uk/publications/rapid-rise-generative-ai (accessed: 23 September 2026).

Juelich, A. (2026) "God Has Helped Us, and So Will AI": How the Terrorist Group Boko Haram Uses Frontier AI. Frontier AI Working Paper Series No. 1/2026. Cambridge Programme on AI Science & Policy, University of Cambridge. Available at: Cambridge CASP report (accessed: 23 September 2026).

Koplin, J.J. (2023) 'Dual-use implications of AI text generation', Ethics and Information Technology, 25, 32. Available at: https://link.springer.com/article/10.1007/s10676-023-09703-z (accessed: 23 September 2026).

Lakomy, M. (2023) 'Artificial Intelligence as a Terrorism Enabler? Understanding the Potential Impact of Chatbots and Image Generators on Online Terrorist Activities', Studies in Conflict & Terrorism. doi: 10.1080/1057610X.2023.2259195. Available at: Taylor & Francis (accessed: 23 September 2026).

Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., Wang, K. and Liu, Y. (2023) 'Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study', arXiv. Available at: https://arxiv.org/abs/2305.13860 (accessed: 23 September 2026).

Molas, B. and Lopes, H. (2024) "Say It's Only Fictional": How the Far-Right is Jailbreaking AI and What Can Be Done About It. ICCT Report. The Hague: International Centre for Counter-Terrorism. doi: 10.19165/2024.5816.

Nasr, M., Carlini, N., Sitawarin, C., Schulhoff, S.V., Hayes, J., Ilie, M., Pluto, J., Song, S., Chaudhari, H., Shumailov, I., Guha Thakurta, A., Xiao, K.Y., Terzis, A. and Tramèr, F. (2026) 'The Attacker Moves Second: Stronger Adaptive Attacks Bypass Defenses Against LLM Jailbreaks and Prompt Injections', 35th USENIX Security Symposium, pp. 1467–1486. Available at: USENIX publication page (accessed: 23 September 2026).

National Institute of Standards and Technology (NIST) (2024a) Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. U.S. Department of Commerce. Available at: NIST publication page (accessed: 23 September 2026).

National Institute of Standards and Technology (NIST) (2024b) U.S. Secretary of Commerce Gina Raimondo Releases Strategic Vision on AI Safety, Announces Plan for Global Cooperation Among AI Safety Institutes. NIST, 21 May. Available at: NIST webpage (accessed: 23 September 2026).

OpenAI (2023) OpenAI's Approach to Frontier Risk. OpenAI, 26 October. Available at: OpenAI webpage (accessed: 23 September 2026).

OpenAI (2024) Improving Model Safety Behavior with Rule-Based Rewards. OpenAI, 24 July. Available at: OpenAI webpage (accessed: 23 September 2026).

Peigné-Lefebvre, P., Feuillade-Montixi, Q., David, T. and Miailhe, N. (2025) LLM Robustness Leaderboard v1: Technical Report. arXiv:2508.06296. Available at: https://arxiv.org/abs/2508.06296 (accessed: 23 September 2026).

Peng, A., Michael, J., Sleight, H., Perez, E. and Sharma, M. (2024) Rapid Response: Mitigating LLM Jailbreaks with a Few Examples. arXiv:2411.07494. Available at: https://arxiv.org/abs/2411.07494 (accessed: 23 September 2026).

Phuong, M., Aitchison, M., Catt, E., Cogan, S., Kaskasoli, A., Krakovna, V., Lindner, D., Rahtz, M., Assael, Y., Hodkinson, S., Howard, H., Lieberum, T., Kumar, R., Abi Raad, M., Webson, A., Ho, L., Lin, S., Farquhar, S., Hutter, M., Delétang, G., Ruoss, A., El-Sayed, S., Brown, S., Dragan, A.D., Shah, R., Dafoe, A. and Shevlane, T. (2024) Evaluating Frontier Models for Dangerous Capabilities. arXiv:2403.13793. Available at: https://arxiv.org/abs/2403.13793 (accessed: 23 September 2026).

Shang, Z. and Wei, W. (2025) 'Evolving Security in LLMs: A Study of Jailbreak Attacks and Defenses', arXiv. doi: 10.48550/arXiv.2504.02080. Available at: https://arxiv.org/abs/2504.02080 (accessed: 23 September 2026).

Tech Against Terrorism (2023a) Early Terrorist Adoption of Generative AI. Tech Against Terrorism, 8 November. Available at: Tech Against Terrorism (accessed: 23 September 2026).

Tech Against Terrorism (2023b) Terrorist Use of Generative AI: Contextualising the Threat and Establishing Frameworks to Prevent the Terrorist and Violent Extremist Exploitation of Generative AI. Available at: Tech Against Terrorism (accessed: 23 September 2026).

Wei, A., Haghtalab, N. and Steinhardt, J. (2023) Jailbroken: How Does LLM Safety Training Fail? arXiv:2307.02483. Available at: https://arxiv.org/abs/2307.02483 (accessed: 23 September 2026).

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J.Z. and Fredrikson, M. (2023) Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv:2307.15043. Available at: https://arxiv.org/abs/2307.15043 (accessed: 23 September 2026).

Author profiles

Dr. Friedhelm Hartwig

Dr Friedhelm Hartwig works as a research associate at modus|zad in the field of Islamism and has been carrying out monitoring projects for the past eight years, the results of which inform professionals in the fields of prevention and civic education about current extremist activities on social media. He studied Islamic Studies, Arabic and Protestant Theology and conducted research at the Leibniz Centre for Modern Oriental Studies in Berlin.

Kaiya M. Reisch

Kaiya M. Reisch works at BAG RelEx, where her work focuses primarily on how generative AI is used for extremist messaging on social media and what opportunities it offers for prevention. Kaiya M. Reisch is a cultural anthropologist (MA, University of Bonn and Tel Aviv University). Her research focuses on collective identity constructions in social discourses and conflicts, as well as how these evolve, including in the context of radicalisation processes.

Heron Lopes

Heron Lopes is a research associate at Stop Hate Brazil and an OSINT Investigator in the private sector. His research focuses on political violence, insurgencies, and the misuse of technology for extremist purposes. He has worked with leading international research institutions, including the International Centre for Counter-Terrorism (ICCT) in the Netherlands, the Uppsala Conflict Data Programme (UCDP) in Sweden, and Leiden University in the Netherlands.

Legal notice

KN:IX connect
Network for Prevention of Islamism and the Promotion of Democracy
Email: info@kn-ix.de
Web: www.kn-ix.de

Publisher:

ufuq.de
Dudenstr. 6
10965 Berlin

Contact Person:

Götz Nordbruch
goetz.nordbruch@ufuq.de

Editing:

Sakina Abushi, Elena Jung, Götz Nordbruch

KN:IX connect | Network for Prevention of Islamism and the Promotion of Democracy was launched by IFAK e. V., ufuq.de, modus|zad, and BAG RelEx to promote networking, the exchange of experiences, and professional development in the field of Islamism prevention.

KN:IX connect aims to sustainably strengthen the prevention of Islamism at the federal level and to promote networking among stakeholders in the field of Islamism prevention. In addition, we pool our expertise and continue to develop our professional knowledge in order to make knowledge and methods accessible to a broad audience.

With KN:IX connect we target, among others, professionals in the field of Islamism prevention and related areas, as well as opinion leaders and professionals from schools, academia, government, and (security) agencies, as well as state democracy centers and partnerships for democracy. In addition, we are available to politicians and the media as points of contact.

KN:IX connect is funded under the federal program “Live democracy!”.

Trend reports are the official publication format of KN:IX connect. Their purpose is to identify trends and challenges in the field of preventing Islamism at an early stage and to analyze them systematically in order to make the insights gained accessible to stakeholders and to promote the exchange between research and practice.

This publication does not represent the views of the BMBFSFJ, the BAFzA, or any other funding agencies.
The authors of this publication are responsible for its content.

ISSN 3053-626X

Ein Verbund von

BAG RelEx | modus | zad | ufuq.de | IFAK e.V.

Gefördert vom Bundesministerium für Bildung, Familie, Senioren, Frauen und Jugend
im Rahmen des Bundesprogramms Demokratie leben!