KN:IX connect Verbund Islamismusprävention und Demokratieförderung

Trendreport #7

KI extrem:

Künstliche Intelligenz und ihre Auswirkungen auf die Islamismusprävention

Autor*innen:
Dr. Friedhelm Hartwig, Kaiya Reisch,
Heron Lopes

Inhalt

Vorwort 4

KI-Imame, Chatbots und Apps – Herausforderungen für Monitoring, Prävention und politische Bildung *(Autor: Dr. Friedhelm Hartwig)* **5**

Wie wird KI von relevanten Akteur*innen genutzt? 6
Welche Haltungen von Creator*innen und User*innen sind sichtbar 6
Schlüsse für Monitorings, Prävention und politische Bildung 8
Fazit 9

Dschihad mit KI: Wie islamistische Akteur*innen generative künstliche Intelligenz nutzen *(Autorin: Kaiya M. Reisch)* **12**

Von Avataren bis Algorithmen: Anwendungsfelder im Überblick 13
Wenn Guardrails fallen: Risiken von Zuschreibung, Regulierung und Nutzung 15
Ein Wettlauf gegen entfernbare Schutzmechanismen 15

Jenseits des Jailbreakings: Stärkung der Resilienz gegenüber KI-gestütztem Extremismus *(Autor: Heron Lopes)* **18**

Bekannte Anwendungsfälle von KI 19
Jailbreaking 21
Fazit 23

Autor*innenprofile 26

Vorwort

Digitale Räume und ihre entsprechenden Zugangs-, Verbreitungs- und Gruppendynamiken können in Radikalisierungsprozessen eine entscheidende Rolle spielen. Umso wichtiger ist es für Präventionsarbeit, die Themen, Trends und technischen Entwicklungen der digitalen Welt eng zu verfolgen und ihre Wirkmacht für ihre Zielgruppen zu begreifen.

Die Möglichkeiten generativer künstlicher Intelligenz (KI) lassen bereits jetzt tiefgreifende gesellschaftliche Veränderungen sowie neue Herausforderungen erkennen. Islamistische Akteur*innen – bereits in der Vergangenheit *early adopters* neuer Technologien – greifen immer häufiger auf KI-gestützte Anwendungen zurück, um die Sichtbarkeit und Reichweite ihrer Inhalte zu erhöhen.

Dieser Trendreport beleuchtet unterschiedliche Dimensionen der Nutzung von generativer KI durch islamistische Akteur*innen. So wird dargelegt, wie sie ihre Arbeit systematisch mit KI-Anwendungen professionalisieren, Reichweiten signifikant erhöhen und ihre Medienarbeit entsprechend KI-Metriken ausrichten. Dabei spielt auch das ambivalente Verhältnis zur technischen Neuerung eine Rolle.

Nicht zuletzt bearbeitet der Trendreport anhand einschlägiger Beispiele den Einsatz von KI für terroristische Zwecke und spannt somit den Bogen von Akteur*innen der (digitalen) Grauzone bis zu gewaltbereiten Extremist*innen.

In den drei Beiträgen wird zudem diskutiert, welche Implikationen und Handlungsbedarfe sich für Akteur*innen aus Zivilgesellschaft, Sicherheitsbehörden, Politik und Verwaltung ergeben.

KI-Imame, Chatbots und Apps — Herausforderungen für Monitoring, Prävention und politische Bildung

Autor:
Dr. Friedhelm Hartwig

*Von künstlicher Intelligenz generierte Inhalte und Tools eröffnen islamistischen Akteur*innen vielfältige Optionen und erscheinen in ihren gesellschaftlichen Auswirkungen noch tiefgreifender als die digitale Revolution durch Internet und Social Media (Lang 2025; Baele 2024). Dieser Artikel fasst die Erkenntnisse des Monitorings von modus|zad zur Nutzung von KI durch islamistische Akteur*innen und User*innen seit 2023 zusammen. Dabei werden neben neuen Formaten und ihrer vereinfachten Verbreitung auch kontroverse Haltungen sowie eine problematische Naivität von User*innen in der Nutzung von KI sichtbar. Der Artikel beschreibt daher auch notwendige Modifizierungen des Monitorings sowie Herausforderungen für Prävention und politische Bildung.*

Wie wird KI von relevanten Akteur*innen genutzt?

Seit 2023 fällt im Monitoring eine zunehmende Nutzung KI-generierter Visualisierungen auf, mit denen islamistische Akteur*innen versuchen, Aufmerksamkeit für ihre Inhalte zu erzielen (Hartwig et al. 2023, 10ff.). Häufig veranschaulichen sie in Bildern oder Videosequenzen Fantasien über heroische Ereignisse der islamischen Frühzeit, das Leben der Propheten, die Wohltaten im Paradies oder Dystopien der Apokalypse. Videos zu koranischen Wesen wie Engel, Dschinns oder dem Dajjal, einem zerstörerischen Vorboten der Apokalypse, erreichen Aufrufzahlen in fünfstelligen Bereichen. Dabei entstehen Bilderwelten, die an Gaming, Fantasy oder idealisierende und romantisierende Darstellungen des Orientalismus erinnern (KN:IX plus, Viewing Session 2). Zumeist sind es anonymisierte Kanäle aus den USA und Südostasien, die ein weltweites mehrsprachiges Angebot aufbauen („Religious Storytelling", „Deen Stories TV"). Es gibt jedoch auch europäische Akteur*innen, zum Beispiel aus Bosnien-Herzegowina („KI Glaubensreise") oder aus Malta („Religious Story TV German").

KI unterstützt dabei nicht nur die Visualisierung und Textgestaltung im Storytelling-Format, sondern auch die Übersetzung in zahlreiche Sprachen. Plattformen fördern diesen Trend durch ihre automatische Synchronisierungsfunktion. So erscheinen auch einzelne ausländische Akteur*innen mittlerweile mit deutschsprachigen Videos („Majed Mahmoud, Brother Aqib"). Weit verbreitet sind auch KI-generierte diffamierende Verfremdungen von Persönlichkeiten, die zu Feinden des Islams erklärt werden, sowie antimodernistische Argumentationen („Religious Story TV German"). Auch die Zahl antisemitischer Darstellungen hat durch die einfach zu bedienenden KI-Bildgeneratoren erheblich zugenommen (Brodersen et al. 2026).

Welche Haltungen von Creator*innen und User*innen sind sichtbar?

Das salafistische Spektrum in Deutschland

Aufschlussreich ist ein Video des salafistischen Akteurs Abul Baraa. In diesem Video präsentiert er eine Antwort von ChatGPT, die ihn angeblich als glaubwürdigen Prediger bestätigt, der insbesondere denjenigen eine Orientierung geben kann, die sich neu mit dem Islam beschäftigen („abulbaraatube", 07.10.2025). Sehr wahrscheinlich handelt es sich in diesem Fall um eine Manipulation. Aufschlussreich ist auch die Analyse der Kommentare. Diese fallen überwiegend positiv und bestätigend aus. So versichern auch viele Kommentierende, dass ChatGPT ein verlässliches Tool sei, um Informationen über den Islam zu gewinnen. Nur vereinzelt findet man dazu kritische Stimmen. Die überwiegende Mehrheit der Kommentierenden nutzt KI-Chatbots nach eigener Darstellung unkritisch als Entscheidungsinstanz und Wissensquelle.

Eine umfassendere Untersuchung der Inhalte deutschsprachiger, salafistischer Akteur*innen zur Anwendung von KI zeigt jedoch auch ein ambivalentes Verhältnis. Einerseits wird KI gerne als Bestätigung ihrer Botschaften und als Hilfsmittel genutzt. Andererseits warnen Akteure wie Abul Baraa oder Pierre Vogel in ihren Videos vor Fehlern und Gefahren, die sie in der Anwendung von KI erleben. So betonen sie, dass man Wissen benötige, um KI anzuwenden. Fehler verzeichnen sie beispielsweise bei der Verknüpfung islamischer Quellentexte mit deren Autoren und Übersetzungen aus dem Arabischen. Eine ungeprüfte Akzeptanz von KI-generierten Antworten

habe zu schweren Streitereien über Glaubensfragen in Moscheen geführt. Letztendlich, so die Botschaft der meisten untersuchten salafistischen Akteur*innen, könne KI nicht das Wissen islamischer Gelehrter ersetzen (KN:IX plus, Viewing Session 2).

Eine umfassende Kritik an Social Media und KI verbreitet die salafistische Akteurin Hanna Hansen (KN:IX plus, Viewing Session 2), obwohl digitale Räume auch für ihre Popularität die wichtigsten Bühnen sind. Social Media verführe zur Zeitverschwendung und lenke von einem bewussten religiösen Leben ab, es handele sich dabei um ein Werkzeug westlicher Ideologie und Unmoral. Dagegen biete, so Hanna Hansen, der Islam mit seinen vorausschauenden, zeitlosen Regeln insbesondere für Frauen wirksamen Schutz und Prävention. Der Fall Collien Fernandez zeige die „digitale Erniedrigung, Identitätsmissbrauch, Einsatz von KI gegen eine Frau. (...) Die Frau wird angeschaut, bewertet, benutzt und heute sogar digital manipuliert und entblößt" („hannahansenofficial", 23.03.2026). Auch vor dem leichtfertigen Gebrauch von ChatGPT warnt sie. Die KI liefere zwar Informationen, diese können jedoch falsch sein und müssten überprüft werden. Besser sei es, die Quellen selbst zu konsultieren („Hanna Hansen", 29.04.2025).

Frühe Experimente von Anas Islam und Kritik von Marcel Krass

Neben salafistischen Akteur*innen beschäftigen sich auch deutschsprachige islamische Speaker*innen, die sich von der salafistischen Szene distanzieren, mit KI (Hartwig et al. 2023,10ff.). Anas Islam war 2023 einer der ersten, der mit KI-Bildgeneratoren und ChatGPT experimentierte. Auch sein thematisches Spektrum erfasst die zuvor genannten religiösen Themen, wie Vorstellungen von der islamischen Apokalypse, dem Jüngsten Gericht, Hölle und Paradies („Anas Islam", 01.04.2025). Seine Experimente mit ChatGPT erzeugen im deutschsprachigen Publikum überdurchschnittlich hohe Aufrufzahlen im fünf- bis sechsstelligen Bereich. Die Themen und ihre medialen Umsetzungen mittels KI scheinen somit ein tiefes Bedürfnis im Publikum zu treffen. Anas Islam nutzt ChatGPT zudem als Tool für religiöse Gesprächsthemen, beispielsweise zu einer Beweisführung der Existenz Gottes („Anas Islam", 29.03.2023, 26.06.2024). Man findet zahlreiche ähnliche Beispiele in international rezipierten Kanälen, wie zum Beispiel einen Disput zwischen einem KI-generierten christlichen und einem muslimischen Avatar („TowardsEternityDeutsch", 02.12.2025).

Aufschlussreich sind auch hier sowohl die Haltung des Creators als auch die Reaktionen der Kommentierenden. So bewertet Anas Islam in seinen Videoinszenierungen den KI-Chatbot als bedeutsames, glaubwürdiges Werkzeug für existenzielle, religiöse Fragestellungen. In einem Video steigert Anas Islam die Inszenierung dahingehend, dass ChatGPT angeblich sogar die Existenz Gottes bestätige. Damit suggeriert Anas Islam seinem Publikum, dass man ChatGPT als Entscheidungsinstanz in tiefgreifenden Glaubensfragen nutzen könne.

Überdurchschnittlich hoch ist auch die Zahl der Kommentare zu den KI-Inszenierungen von Anas Islam. Die Kommentare sind sehr emotional und stimmen den Aussagen der Videos überwiegend zu. Zahlreiche Kommentierende wünschen sich weitere Visualisierungen, Diskussionen mit ChatGPT und KI-generierte Koranauslegungen. Im Gegensatz zu Anas Islam kritisiert Marcel Krass auf dem Kanal „Föderale Islamische Union (FIU)" die Entwicklungen zu KI-generierten Inhalten und Apps („FIU", 07.03.2026). Er räumt zwar ein, dass KI eine Hilfe sein könne, sie sei aber nicht neutral. Er warnt daher sein Publikum vor einem unkritischen Umgang. Mit Sorge beobachte er einen weit verbreiteten Gebrauch von KI als Entscheidungsinstanz in Glaubensfragen, mahnt zu eigenverantwortlichem Handeln sowie einer Orientierung an islamischen Gemeinden.

Im Vergleich zu Anas Islam erreichen die kritischen Videos zu KI-Chatbots von Marcel Krass allerdings nur wenige Aufrufe und Kommentare. Diese stimmen der Kritik von Marcel Krass weitestgehend zu. KI sei nur ein Geschäft. Einige befürchten sogar, man befände sich nun in einer Vorstufe zur Apokalypse, einer Zeit der Verwirrung und Täuschung.

Die Mehrheit der Akteur*innen und ihr Publikum scheinen jedoch nahezu kritiklos die Nutzung von KI-Systemen als quasi menschliche Berater oder

KI-Imame in ihren Alltag übernommen zu haben. Influencer*innen suggerieren ihrem Publikum, die Chatbots hätten überlegenes Wissen, Überzeugungen, Gefühle oder sogar einen Glauben. Zudem sind die KI-Systeme so programmiert, dass sie emotionale Bindungen an sie initiieren und steigern. In der Forschung diskutiert man die Zuschreibung von menschlichen Eigenschaften an maschinelle Prozesse und die daraus folgende mangelnde Unterscheidung zwischen einer simulierten und realen Subjektivität unter Arbeitsbegriffen wie digitaler Animismus (Fuchs, 2025) oder digitale Religion (Prohl, 2026). Notwendig ist - so die Forscher*innen - eine aktive Distanzierung gegenüber einer zunehmenden digitalen Vortäuschung von menschlichen Eigenschaften durch KI-Systeme.

Schlüsse für Monitorings, Prävention und politische Bildung

Deepfake und Desinformation

Spätestens nach dem Massaker vom 7. Oktober 2023 wurde deutlich, dass die Flut an KI-generierten Deepfakes und Desinformationen mit den zur Verfügung stehenden begrenzten Mitteln nicht mehr bewältigt werden kann. Die Moderation von Plattformen zeigt zudem wenig Effizienz. Ressourcen für einen hinreichend guten Faktencheck, die insbesondere in Krisenzeiten schnell aufgestockt werden können, wären hier hilfreich. Auch die Bereitstellung von Open-Source-Tools könnte dabei unterstützen, das Internet und Social Media zu zuverlässigeren Informationsquellen und hochwertigeren Kommunikationsräumen zu machen.

Größere Datenmengen, beschleunigte Produktion, höhere Qualität

KI-generierter Content, automatische Synchronisation und die einfache Anwendung mit einer immer besser werdenden Qualität führen zu einer Vervielfältigung der Kanäle und dem Auftreten neuer Akteur*innen. Monitorings müssen sich daher auf die Analyse noch größerer Datenmengen sowie das Auftreten neuer Akteur*innen, die für die Prävention relevant sein könnten, einstellen. Entsprechend flexibel sollten sich daher auch Fachkräfte der Prävention und politischen Bildung aufstellen.

Transnationale Überlagerungen der Akteur*innen

Das bislang übersichtliche Feld deutschsprachiger islamistischer Kanäle wird durch die wachsende Zahl transnationaler Kanäle in deutscher Sprache unübersichtlicher. International agierende Kanäle verfügen häufig über deutlich höhere Aufruf- und Abonnent*innenzahlen als deutschsprachige Kanäle. Inwieweit diese Kanäle für deutschsprachige User*innen bedeutsam werden oder sogar zu einer Verdrängung des etablierten Spektrums führen, müssen zukünftige Studien zeigen.

Chatbots und KI-Apps — neue Aufgaben für Monitoring und Prävention

Ein Feld, das dringend wissenschaftlich analysiert werden müsste, ist der schnell wachsende Markt islamischer Chatbots. Inwieweit diese seriöse und zuverlässige Angebote darstellen oder auch als Mittel zur Verbreitung islamistischer Propaganda und Rekrutierung dienen könnten, wird ein neues Forschungsfeld im Monitoring eröffnen. Spannend sind zudem Bemühungen islamischer Akteur*innen, eine eigene KI-Souveränität zu behaupten und Angebote zu entwickeln, die den jeweiligen Idealen und Visionen ihrer Entwickler*innen entsprechen. Auch hier werden sehr wahrscheinlich neue Akteur*innen ins Spiel kommen, die bislang in der Prävention und politischen Bildung unbekannt sind.

Fazit

Werden KI-Imame, Chatbots und Apps ein weiteres effektives Mittel islamistischer Akteur*innen werden? Verdrängen diese möglicherweise sogar als allzeit verfügbare digitale Instanzen reale Influencer*innen oder gar Imame und Moscheegemeinden? Wie wird sich das Verhältnis von Glauben, Wissen und Autorität durch KI-generierte Inhalte generell verändern? Lassen sich einige Kommentierende bereits als Anhänger*innen einer digitalen Variante des Islams beschreiben?

Betrachtet man die Ebene der Nutzer*innen, kann man sagen, dass die überwiegende Mehrheit der User*innen Chatbots und KI-generierte Inhalte äußerst unkritisch zu nutzen scheint. KI wird als umfassende religiöse Entscheidungsinstanz akzeptiert und als Wissensquelle genutzt, ohne die Resultate zu prüfen. Die Kommentare zeigen zudem eine hohe emotionale Resonanz. Viele User*innen reagieren mit großer Zustimmung und Dankbarkeit, fühlen sich im Glauben bestärkt und wünschen sich mehr Content in dieser Form. Zahlreiche Beispiele aus dem Monitoring belegen, dass der KI eine Aura von überhöhten menschlichen Fähigkeiten und religiösen Kompetenzen zugeschrieben wird. Ein Trend zum digitalen Animismus ist daher deutlich erkennbar. Dieser digitale Animismus wird sehr wahrscheinlich eine neue Herausforderung im zukünftigen Aufgabenbereich von Prävention und Medienkompetenz sein. Notwendig sind insbesondere Kompetenzen einer kritischen Reflektion darüber, wie digitale Kulturen und KI-generierte Angebote die Gesellschaft und das eigene Denken, Fühlen und Handeln beeinflussen. Notwendig ist auch ein kritisches Bewusstsein gegenüber denjenigen Angeboten, die KI mit religiöser Rhetorik als vermeintlichen Heilsbringer, allumfassende Problemlösung und damit als digitalen Götzen propagieren.

Literaturverzeichnis:

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 (abgerufen am 11.08.2026).

Brodersen, W., Schmidt, M., Fielitz, M. 2026. „Portionierter Judenhass. Eine Vermessung des Antisemitismus auf TikTok." *In Machine Against the Rage 9* (2026). https://machine-vs-rage.net/ausgabe-9/portionierter-judenhass/ (abgerufen am 11.08.2026).

Bundesministerium des Inneren. 2026: *Verfassungsschutzbericht 2025.* https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/sicherheit/BMI26012-vsb2025.pdf?__blob=publicationFile&v=7 (abgerufen am 11.08.2026).

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 (abgerufen am 15.12.2025).

Hartwig, F., Seelig, J., Buğur, D. 2023. *Monitoring der Peripherie des religiös begründeten Extremismus (PrE). Quartalsbericht 2/2023.* https://www.bpb.de/system/files/dokument_pdf/bpb_Basismonitoring_QB2_2023.pdf?download=1 (abgerufen am 11.08.2026).

Hartwig, F. 2024. „Islamismus in den sozialen Medien: Eskalationsspiralen auf TikTok und You-Tube." [DiS]Like: *Soziale Medien zwischen Zusammenhalt und Polarisierung* 3: 79-101. Hentrich & Hentrich Verlag Leipzig.

KN:IXplus. 2026. *Viewing Session 2.* https://knixplus.modus-zad.de/ (abgerufen am 11.08.2026).

Lang, I. 2025. *Künstliche Intelligenz und politischer Extremismus. Ein Überblick.* Springer VS. https://link.springer.com/book/10.1007/978-3-658-49024-9 (abgerufen am 11.08.2026).

Modus|zad o. J. *Monitorings und Trendanalysen.* https://modus-zad.de/schwerpunkte/monitorings-trendanalysen/ (abgerufen am 11.08.2026).

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 (abgerufen am 03.04.2026).

Social Media Quellen:

„Brother Aqib“, https://www.youtube.com/@BrotherAqibOfficial (abgerufen am 23.09.2026).

„Deen Stories TV“, https://www.youtube.com/@DeenStoriesTV (abgerufen am 23.09.2026).

„KI Glaubensreise“, https://www.youtube.com/@KI-Glaubensreise (abgerufen am 23.09.2026).

„Majed Mahmoud“, https://www.youtube.com/@MajedMahmoud (abgerufen am 23.09.2026).

„Religious Story TV German“, https://www.youtube.com/@ReligiousStoryTV (abgerufen am 23.09.2026).

„Religious Storytelling“, https://www.youtube.com/@ReligiousStorytellingg (abgerufen am 23.09.2026).

„Anas Islam“. 29.03.2023. *ChatGPT bestätigt ALLAHs EXISTENZ! KI widerlegt Atheismus!* https://www.youtube.com/watch?v=3cWMrwBPGH4 (abgerufen am 11.08.2026).

„Anas Islam“. 26.06.2024. *Evolution spricht für ALLAHs Existenz!! Diskussion mit sprechender KI (ChatGPT)*. https://www.youtube.com/watch?v=GrKuxCJgsbs (abgerufen am 11.08.2026).

„Anas Islam“. 01.04.2025. *ChatGPT (KI) zeichnet Hölle und Paradies*. https://www.youtube.com/watch?v=q7O6CnqlBbI (abgerufen am 11.08.2026).

„abulbaraatube“, 07.10.2025. *Abul Baraa fragt Chatgpt nach Abul Baraa Antwort war überraschend!* https://www.youtube.com/watch?v=gTaLF9QpmC4 (abgerufen am 11.08.2026).

„FIU (Föderale Islamische Union), Marcel Krass“. 07.03.2026. *KI-Halluzinationen und religiöse Fakten – Marcel Krass.* https://www.youtube.com/watch?v=2cMBZAcg4Wc (abgerufen am 11.08.2026).

„FIU (Föderale Islamische Union), Marcel Krass“. 11. Aug. 2025. *Künstliche Scheichs: Die große Täuschung beginnt – Marcel Krass.* https://www.youtube.com/watch?v=g0QYoJcNag0 (abgerufen am 11.08.2026).

„hannahansenofficial“. 23.03.2026. *Digitale Erniedrigung im Westen. Schutz & Würde im #Islam! #collienfernandez #deefakes.* https://www.youtube.com/watch?v=-u-RMZtUrHQ (abgerufen am 11.08.2026).

„Hanna Hansen“. 29.04.2026. *Vortrag Hanna Hansen in Medina.* BildungsPaket (21.04.25). https://www.youtube.com/watch?v=rY-UEmLymgg (abgerufen am 11.08.2026).

„TowardsEternityDeutsch“. 02.12.2025. *Ist Jesus Gott oder ein Prophet – Die große KI-Debatte.* https://www.youtube.com/watch?v=3Y8Nh6JGGMY (abgerufen am 11.08.2026).

„ummatic“. 08.11.2025. *Cyber-Umma, Algorithmic Hegemony, and Digital Decolonization.* https://www.youtube.com/watch?v=eWp_6ifRGDs (abgerufen am 11.08.2026).

„ummatic“. 27.12.2025. *Towards an Ummatic Vision for Advanced AI.* https://www.youtube.com/watch?v=yZdIAZwkyDY (abgerufen am 11.08.2026).

Dschihad mit KI: Wie islamistische Akteur*innen generative künstliche Intelligenz nutzen

Autorin:
Kaiya M. Reisch

*Islamistische Akteur*innen setzen generative künstliche Intelligenz zunehmend systematisch ein: von KI-Sprecher*innen und Deepfake-Propagandavideos über automatisierte Rekrutierungs-Chatbots bis hin zu KI-gestützter Zielgruppenanalyse. Dieser Beitrag ordnet die Nutzung von KI zunächst inhaltlich ein und zeigt anhand aktueller Beispiele, wie islamistische Gruppen generative KI in ihre Medienarbeit integrieren und welche regulatorischen und rechtlichen Unwägbarkeiten sich aktuell stellen. Der Beitrag plädiert für eine engere Zusammenarbeit zwischen Tech-Unternehmen, Sicherheitsbehörden und zivilgesellschaftlicher Präventionsarbeit, die durch ihren direkten Zugang zu Jugendlichen eine zentrale, bislang zu wenig genutzte Rolle einnehmen kann.*

Islamistische Bewegungen gelten in der Forschung als *early adopters* neuer Technologien. Schon die Nutzung von Satellitenfernsehen, Foren, Verschlüsselungs-Apps und maschineller Übersetzung folgte diesem Muster: Neue Werkzeuge werden früh erprobt, um Reichweite, Anonymität und Professionalität der eigenen Medienproduktion zu steigern (Scheeres & Tabti 2026). Generative KI passt strukturell in dieses Profil: Sie ist global verfügbar, kostengünstig und senkt Zugangshürden drastisch. Das ist insbesondere für Netzwerke, die auf Dezentralisierung und mediale Präsenz über Ländergrenzen hinweg angewiesen sind, ein Vorteil (Lang 2026). Der sogenannte Islamische Staat (IS) und ihm nahestehende Akteure (etwa der Islamische Staat Khorasan Provinz, ISKP) haben ihre mediale Infrastruktur über die Jahre professionell ausgebaut; KI fügt sich hier als weiteres Produktionswerkzeug ein, nicht als Bruch mit bestehender Praxis.

Gleichzeitig ist das Verhältnis zu dieser Technik durchaus ambivalent: Innerhalb islamistischer Onlinecommunitys existiert neben der frühen Übernahme auch eine ausgeprägte KI-Skepsis, etwa wegen theologischer Vorbehalte gegenüber bildgenerierender KI oder wegen eines grundsätzlichen Misstrauens gegenüber als „westlich" wahrgenommener Technologie (Scheeres & Tabti 2026). Die Nutzung ist also kein Selbstläufer, sondern verläuft uneinheitlich. Während offizielle Medienstellen wie die des IS professionell und mit klarer strategischer Ausrichtung arbeiten, experimentieren andere Milieus eher unsystematisch, aus Neugier oder um sich innerhalb der eigenen Szene zu profilieren. Das Verhältnis von Islamismus und KI reicht insofern von hochorganisierter Propagandaproduktion auf der einen Seite bis hin zu diffuser, individueller Aneignung dieser Werkzeuge auf der anderen.

Von Avataren bis Algorithmen: Anwendungsfelder im Überblick

Am sichtbarsten wird der Einsatz von KI bislang bei der Content-Produktion. Nach dem Anschlag auf die Crocus City Hall im Moskauer Vorort Krasnogorsk im März 2024 setzte der IS KI-generierte Avatar-Sprecher für eigene Nachrichtenformate ein (Lang 2026). Auffällig ist dabei, wie gezielt die Produktion mittlerweile auf jugendliche Zielgruppen zugeschnitten ist: So produzierten Sympathisant*innen von al-Qaida und dem IS beispielsweise Deepfake-Musikvideos, in denen bekannte Figuren wie Spongebob Schwammkopf oder populäre YouTuber wie MrBeast islamistische Kampfgesänge vortragen (Lang 2026; Katz 2024). Solche Formate funktionieren gerade deshalb, weil sie bestehende popkulturelle Vertrautheit nutzen, um ideologische Inhalte niedrigschwellig und emotional aufgeladen zu vermitteln. Islamistische Gruppen setzen ihre KI-Tools dabei plattformspezifisch ein: Auf TikTok dominieren kurze, ästhetisch ansprechende Formate, die gezielt Vertrautheit simulieren (Lang 2026; Siegel 2024).

Ergänzt wird dieses ästhetische Element vielfach durch ein strategisches: Das Islamic Media Cooperation Council, eine al-Qaida-nahe Mediengruppe, veröffentlichte 2024 eine arabischsprachige Anleitung zur Nutzung von ChatGPT für Propagandazwecke, und der IS selbst publizierte bereits 2023 einen eigenen Leitfaden zur „sicheren" Nutzung generativer KI (Katz 2024; Minniti 2025). Es geht also nicht mehr um ein Ausprobieren und Erproben, sondern um eine strategisch ausgerichtete Wissensvermittlung mit dem Ziel, neue Mitglieder dazu zu befähigen, KI-Werkzeuge routinemäßig in der eigenen Medienarbeit zu nutzen.

Über die reine Inhaltserstellung hinaus gewinnt zudem die automatisierte Rekrutierung neuer Mitglieder an Bedeutung. So nutzt der ISKP beispielsweise KI, um Propagandabulletins mit Deepfake-gestützten Bildern und synthetischen Stimmen zu produzieren, und setzt zunehmend automatisierte Chatbots ein, die kontinuierlich über mehrere Plattformen hinweg mit potenziellen Rekrut*innen interagieren, deren Verhaltensmuster analysieren und Radikalisierungsimpulse an ideologische Neigungen und individuelle Vulnerabilitäten anpassen (Minniti 2025). Die KI ermöglicht dabei eine Rund-um-die-Uhr-Präsenz und erlaubt parallele Interaktionen mit einem größeren Personenkreis.

So nutzen islamistische Akteur*innen auch KI-generierte Profile, um beispielsweise den Israel-Gaza-Krieg zu instrumentalisieren und islamistische Narrative zu bestärken (jugendschutz.net 2025). Verbreitet ist dabei das Stilmittel der sogenannten Ghiblification: Fotos werden mithilfe von Bildgeneratoren in eine Anime-Ästhetik überführt, um politisch gefärbte Inhalte unauffälliger und für ein jüngeres Publikum ansprechender zu verbreiten. Inhaltserstellung, Identitätskonstruktion und strategische Verbreitung verschränken sich so zu einem Gesamtsystem, das kaum noch durchschaubar ist: Wer online mit einem vermeintlich authentischen Profil interagiert, kann kaum erkennen, ob dahinter eine reale Person oder ein KI-gestütztes Konto steht (jugendschutz.net 2025; Lang 2026).

Dieser Nutzung von KI in der Content-Produktion ist in vielen Fällen eine datenbasierte Zielgruppenanalyse vorgelagert. Sogenannte deskriptive KI wertet das Nutzer*innenverhalten aus, um Personen mit erhöhter Ansprechbarkeit für bestimmte Narrative zu identifizieren. Neben demografischen Merkmalen spielen dabei auch psychologische Faktoren eine Rolle – etwa Zeichen von Frustration, politischer Entfremdung oder ein ausgeprägtes Zugehörigkeitsbedürfnis –, die bei Jugendlichen und anderen vulnerablen Gruppen verstärkt auftreten können (Schroeder et al. 2026).

Die vielfältigen Formen, mit denen islamistische Akteur*innen auf KI zurückgreifen, stehen für die besondere kulturelle Anschlussfähigkeit dieser Strömungen. Eine aktuelle ICCT-Analyse beschreibt, wie der IS seit seiner territorialen Niederlage 2018 eine „virtuelle Kalifat"-Präsenz aufgebaut hat, die gezielt in den digitalen Identitätsräumen der Generation Z operiert: In Gaming-Plattformen wie Roblox und Minecraft kursieren dschihadistisch geprägte Mods und Rollenspielszenarien, während Meme-Formate, Musikgenres wie Phonk (ein Subgenre des Hip-Hop) und plattformtypische Ästhetiken genutzt werden, um ideologische Inhalte in eine für Jugendliche vertraute visuelle und sprachliche Grammatik zu übersetzen (Tahir & Jadoon 2026). Kennzeichnend ist dabei eine „Trichterstrategie": Der Einstieg erfolgt über leicht zugängliche, algorithmisch verstärkte Inhalte auf Mainstreamplattformen, bevor Interessierte schrittweise in stärker ideologisierte, teils verschlüsselte Räume weitergeleitet werden. Diese Beobachtungen zeigen, dass KI-generierte Inhalte nicht isoliert wirken, sondern Teil eines umfassenderen, plattformübergreifenden Ökosystems sind, in dem algorithmische Verstärkung und kulturelle Mimikry ideologische Botschaften strukturell in jugendkulturelle Formate einbetten.

Die technologische Wirkung entfaltet sich im Zusammenspiel dieser KI-gestützten Medienstrategien: Generative Inhaltserstellung, datengetriebene Ansprache und algorithmische Verstärkung greifen ineinander und verstärken sich wechselseitig – ein Mechanismus, der bereits vor der aktuellen KI-Welle als zentraler Verstärker extremistischer Aufmerksamkeitsökonomien beschrieben wurde (Kukaj 2020), aber in der Vergangenheit nur unter Aufwendung großer Ressourcen möglich war. Heute stehen diese Strategien selbst kleinen, dezentral organisierten Gruppen zur Verfügung (Minniti 2025; CIGI 2026). Wie weit diese Institutionalisierung bereits reicht, zeigt der Fall organisierter KI-Nutzung durch islamistisch-dschihadistische Gruppen in Nigeria: So betreiben Nachfolgeorganisationen von Boko Haram eigene „KI-Einheiten", die über VPNs und bezahlte Zugänge kommerzielle Modelle wie ChatGPT, Claude, Gemini, Grok, Meta AI und DeepSeek verwenden (Militant Wire 2026).

Wenn Guardrails fallen: Risiken von Zuschreibung, Regulierung und Nutzung

Während sich die bisher beschriebenen Anwendungsfälle noch relativ direkt beobachten und einzelnen Akteur*innen zuordnen lassen, geht die technische Herausforderung inzwischen darüber hinaus und betrifft zunehmend quelloffene (*open-weight*) Sprachmodelle, deren Schutzmechanismen sich gezielt entfernen lassen. Der entscheidende Unterschied ist folgender: Ein geschlossenes Modell läuft ausschließlich über die Infrastruktur des Anbieters, der Zugriffe sehen, einschränken oder widerrufen kann. Ein offenes Modell wird dagegen als herunterladbare Datei veröffentlicht, die jede Person mit ausreichender Rechenleistung lokal betreiben, verändern und weiterverbreiten kann (Hadley 2026) – frei herunterladbar für jede*n und dabei dem Zugriff und der Regulierung des ursprünglichen Betreibers entzogen.

Für die Präventionsarbeit geht damit eine doppelte Verschiebung einher: Das Risiko verlagert sich von der Frage, ob ein Anbieter seine Modelle absichert, hin zur Frage, ob sich Sicherheitsmechanismen überhaupt dauerhaft in ein Modell verankern lassen. Zudem ist die Größe eines Modells keine echte Hürde mehr: Auch sehr große Modelle, die kaum jemand selbst betreiben kann, werden für alle nutzbar, sobald ein Anbieter mit ausreichender Infrastruktur sie hostet.

Aus diesen Entwicklungen ergeben sich mehrere Unwägbarkeiten, die eng miteinander zusammenhängen. Die erste betrifft die Attribution, also die Zuschreibung der Urheberschaft: Ob ein KI-generierter Output tatsächlich einer extremistischen Gruppe zuzuschreiben ist oder von einzelnen, nicht formell angebundenen Sympathisant*innen stammt, lässt sich oft kaum zweifelsfrei klären. Diese Unschärfe erschwert die juristische Einordnung ebenso wie die Priorisierung begrenzter Ressourcen.

Damit eng verknüpft ist eine Unterscheidung zwischen organisierter, gruppengesteuerter Nutzung – etwa durch offizielle IS-Medienstellen – und einer diffusen, individuellen Radikalisierung, bei der sich Einzelpersonen ohne direkten Gruppenkontakt allein im Austausch mit einem KI-Modell radikalisieren oder in bereits vorhandenen Überzeugungen bestärkt werden. Gerade diese zweite Form entzieht sich klassischer Bedrohungsanalyse, weil sie sich nicht als konkreter Anschlagsplot manifestiert, sondern als schleichender Prozess: Ein Modell, das Zweifel an extremistischen Positionen zerstreut oder die Rolle eines zustimmenden Gesprächspartners einnimmt, kann im Verlauf mehrerer Gespräche hinweg eine starke Wirkung entfalten, obwohl jedes einzelne dieser Gespräche für sich unauffällig bleibt.

Eine dritte Herausforderung ergibt sich aus dem regulatorischen Rückstand Europas gegenüber offenen Modellen. Der Digital Service Act (DSA) und der AI Act adressieren primär geschlossene, plattformgebundene Systeme mit klaren Verantwortlichen (European Commission 2026a, 2026b). Für einmal veröffentlichte offene Modellgewichte fehlt dagegen ein vergleichbarer Kontrollhebel: Weder Plattformpflichten noch nachträgliche Verbote greifen, sobald eine Datei frei zirkuliert. Daraus folgt jedoch kein pauschales Argument gegen offene Modelle. Der eigentliche Graubereich liegt nicht in der Offenheit an sich, sondern darin, dass Sicherheitsmechanismen meist nur als dünne, nachträglich aufgesetzte Schicht existieren, die sich vom Modell trennen lässt, statt untrennbar in dessen Architektur eingebettet zu sein. Unklar bleibt dabei die Frage, wer die Verantwortung für die Nutzung trägt: der oder die Entwickelnde, der Host oder der oder die Nutzende.

Ein Wettlauf gegen entfernbare Schutzmechanismen

Islamistische Akteur*innen setzen KI heute sowohl für die Zielgruppenanalyse als auch für die Inhaltsproduktion und algorithmisch verstärkte Verbreitung ein. Die eigentliche strukturelle Verschiebung liegt jedoch tiefer: Mit sinkenden Zugangshürden und der wachsenden Verbreitung offener, gewichtsbasierter Modelle verlagert sich das Risiko von einzelnen auffälligen Propagandaprodukten hin zu einer

dauerhaft verfügbaren, kaum kontrollierbaren technischen Infrastruktur. Das ist eine Entwicklung, die längst auch außerhalb islamistischer Kontexte als Gefahr für demokratische Öffentlichkeit diskutiert wird (Denzel 2026). Reaktive Moderation, wie sie Sicherheitsbehörden bislang primär betreiben, trifft zunehmend auf ein Problem, das sich nicht mehr an der Plattformoberfläche, sondern bereits auf Ebene der Modellgewichte selbst entscheidet, lange bevor ein Inhalt überhaupt gepostet werden kann.

Wirksame Prävention braucht deshalb ein technisches Verständnis dieser Mechanismen, das das Wissen aus der klassischen Radikalisierungsforschung ergänzt. Fachkräfte sollten den Unterschied zwischen geschlossenen und offenen Modellen, zwischen Jailbreaking und Abliteration sowie zwischen organisierter und individueller Nutzung kennen. Regulatorisch braucht es einen Ansatz, der offene Modelle nicht pauschal verbietet, aber verbindliche Sicherheitsprüfungen vor ihrer Veröffentlichung einfordert und Verantwortung entlang der gesamten Lieferkette sicherstellt.

Weder die Technologieentwicklung, die Arbeit der Sicherheitsbehörden noch eine konsequente Regulierung bieten allein eine Lösung für die beschriebenen Herausforderungen. Entscheidend ist eine enge Zusammenarbeit mit einem vierten Akteur, der in bisherigen Debatten oft zu kurz kommt: die Zivilgesellschaft. Präventionsarbeit, Jugend- und Sozialarbeit sowie Beratungsstellen haben einen direkten, oft niedrigschwelligen Zugang zu Jugendlichen und anderen Personen, die mit entsprechenden Inhalten in Kontakt kommen – einen Zugang, über den weder Tech-Unternehmen noch Sicherheitsbehörden in vergleichbarer Form verfügen. Wirksame Prävention gründet deshalb nicht allein in besseren technischen Schutzmechanismen oder schärferer Regulierung, sondern entsteht dort, wo Tech-Unternehmen, Sicherheitsbehörden und zivilgesellschaftliche Akteur*innen ihr jeweiliges Wissen zusammenführen: technisches Verständnis der Risiken, sicherheitsbehördliche Lageeinschätzung und die pädagogische Nähe zu Zielgruppen. Ein zentraler Hebel dafür ist der nachhaltige Ausbau von KI- und Medienkompetenzbildung als dauerhaft finanzierter Bestandteil von Schule, Jugendarbeit und Erwachsenenbildung, der junge Menschen und ihre Bezugspersonen befähigt, KI-generierte Inhalte, manipulative Ansprache und algorithmische Verstärkungsmechanismen frühzeitig zu erkennen.

Die Frage ist schließlich nicht mehr, ob islamistische Akteur*innen KI nutzen werden, sondern wie schnell Prävention und Regulierung mit einer Technologie Schritt halten, deren Schutzmechanismen sich schneller entfernen lassen, als neue entwickelt werden.

Literaturverzeichnis:

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 (abgerufen am 20.08.2026).

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 (abgerufen am 20.08.2026).

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/ (abgerufen am 20.08.2026).

Jugendschutz.net. 2025. *Jugendschutz im Internet 2025: Risiken und Handlungsbedarf [Jahresbericht].* https://www.jugendschutz.net/fileadmin/daten/publikationen/jahresberichte/jahresbericht_2025.pdf (abgerufen am 20.08.2026).

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/ (abgerufen am 15.08.2026).

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/ (abgerufen am 15.08.2026).

Militant Wire. 2026. *Open-Weight AI Models and Terrorist Exploitation: A Preliminary Assessment.* https://www.militantwire.com/p/open-weight-ai-models-and-terrorist (abgerufen am 21.08.2026).

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/ (abgerufen am 15.08.2026).

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 (abgerufen am 15.08.2026).

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/ (abgerufen am 15.08.2026).

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 (abgerufen am 15.08.2026).

Jenseits des Jailbreakings: Stärkung der Resilienz gegenüber KI-gestütztem Extremismus

Autor:
Heron Lopes

*Generative KI eröffnet terroristischen Akteur*innen neue Möglichkeiten für Propaganda, Rekrutierung sowie die Planung und Durchführung von Anschlägen. Dieser Beitrag untersucht, in welchen Bereichen eine Nutzung von KI durch terroristische Akteur*innen bisher zu beobachten ist, und verweist dabei vor allem auf die Gefahren, die mit einer gezielten Umgehung der Sicherheitsmechanismen von KI-Modellen verbunden sind. Die Risiken sogenannter Jailbreaks machen den Bedarf an weiterer Forschung und die Entwicklung von Strategien im Umgang mit missbräuchlichen KI-Nutzungen durch extremistische Akteur*innen deutlich.*

Die rasante öffentliche Verbreitung generativer KI (GenKI) hat die Art und Weise verändert, wie Menschen Informationen produzieren, nutzen und teilen. Innerhalb weniger Jahre haben sich Systeme, die Texte, Bilder, Audiodateien und Softwarecode erzeugen, die wie von Menschen erstellt wirken, im Alltag durchgesetzt. Sie unterstützen Bildung, wissenschaftliche Forschung, die Betreuung von Kund*innen oder die Entwicklung von Software. Doch wie bei vielen technologischen Innovationen zuvor hat auch die generative KI grundsätzlich zwei Seiten: Dieselben Fähigkeiten, die nützliche Anwendungen ermöglichen, können auch von Akteur*innen ausgenutzt werden, die auf die Schädigung anderer abzielen (Barrett et al. 2023).

Für gewaltbereite Extremist*innen und terroristische Organisationen eröffnen diese Entwicklungen neue Möglichkeiten. Extremistische Organisationen nutzen neue Technologien bereits seit geraumer Zeit für ihre Propaganda, Rekrutierung, Finanzierung und operativen Aktivitäten (Council of Europe 2025). Generative KI hat das Potenzial, diese Aktivitäten zu beschleunigen: Sie reduziert den Zeit- und Kostenaufwand sowie den Bedarf an fachlichem Know-how, der für die Erstellung überzeugender mehrsprachiger Inhalte, die Automatisierung von Routineaufgaben und die Personalisierung von Botschaften für unterschiedliche Zielgruppen erforderlich ist (Tech Against Terrorism 2023b). Zwar gibt es derzeit nur wenige Hinweise darauf, dass generative KI extremistische Operationen grundlegend verändert hat (Molas & Lopes 2024), doch mehren sich die Anzeichen dafür, dass diese Akteur*innen aktiv mit der Technologie experimentieren und sie an ihre Bedürfnisse anpassen (Council of Europe 2025).

In diesem Beitrag wird der Einsatz generativer KI durch extremistische und terroristische Akteur*innen beleuchtet und untersucht, wie diese Technologien genutzt werden, um Propaganda in verschiedenen Formaten – darunter Text, Bilder, Musik und Videos – zu erstellen, zu übersetzen und zu verbreiten. Anhand von Beispielen aus rechtsextremen, akzelerationistischen und salafistisch-dschihadistischen Szenen wird diese Entwicklung von den Anfängen bis hin zu neueren Anwendungen nachgezeichnet. Dabei wird die Veränderung der Kommunikation, Rekrutierung, Radikalisierung und der operativen Fähigkeiten von Extremist*innen durch generative KI herausgearbeitet.

Bekannte Anwendungsfälle von KI

Spätestens seit 2023 belegen Forschungsergebnisse, dass Terrorist*innen KI für die Rekrutierung und die Planung von Anschlägen in verschiedenen ideologischen und operativen Kontexten einsetzen. Terroristische Akteur*innen experimentieren immer wieder mit den neuesten KI-Modellen und KI-gestützten Technologien. Das zeigen unter anderem Erkenntnisse von Tech Against Terrorism und der Universität Cambridge: Demnach haben der sogenannte Islamische Staat und Boko Haram KI-gestützte Anwendungen für Propaganda, Angriffsplanung, Sprengstoffkonstruktion und die Fehlerbehebung bei Waffen erprobt oder gefördert (Tech Against Terrorism 2023b; Juelich 2026). Auch Einzeltäter*innen nutzen Chatbots, um Angriffsfantasien durchzuspielen und Manifeste zu formulieren oder um Anschläge konkret zu planen (Tech Against Terrorism 2025). Aktuelle Berichte veranschaulichen darüber hinaus, wie generative KI in der Vorbereitung terroristischer Gewalt genutzt werden kann (Juelich 2026).

Diese dokumentierten Fälle von KI-Missbrauch basieren auf früheren Forschungsarbeiten zur Prävention und Bekämpfung von gewalttätigem Ex-

tremismus (PCVE), in denen das Potenzial solcher Technologien für die Ausnutzung durch gewalttätige Extremist*innen untersucht wurde, noch bevor systematische Belege für deren Einsatz durch Terrorist*innen vorlagen. Zuerst konzentrierten sich die Forscher*innen darauf, die Anfälligkeit der damals gerade erst auf den Markt gekommenen GenKI-Plattformen gegenüber Missbrauch für gewalttätige extremistische Zwecke zu testen – dabei legten sie den Schwerpunkt zunächst weniger auf tatsächlich dokumentierte Fälle terroristischer Nutzung als vielmehr darauf, potenzielle Nutzungsmöglichkeiten der Technologie zu erkunden (Barrett et al. 2023; Lakomy 2023). So untersuchte Lakomy 2023, ob öffentlich zugängliche Chatbots und Bildgeneratoren terroristische Propaganda produzieren, extremistisches Material abrufen oder Informationen mit Terrorismusbezug bereitstellen könnten. Dabei zeigte sich, dass die Sicherheitsvorkehrungen je nach Plattform erheblich voneinander abwichen und manche Systeme derart manipuliert werden konnten, dass sie schädliche Ergebnisse lieferten. Dieses Beispiel gab einen ersten Hinweis darauf, wie diese Technologie mögliche Hindernisse bei der Erstellung extremistischer Inhalte überwinden könnte. Allerdings handelte es sich um eine experimentelle Bewertung der Modellfähigkeiten und nicht um Beweise dafür, dass terroristische Organisationen diese Werkzeuge bereits systematisch einsetzten.

Diese Unterscheidung zwischen technologischem Potenzial und tatsächlicher Nutzung verlor jedoch zunehmend an Bedeutung, schließlich gab es noch im selben Jahr erste Hinweise auf konkrete Versuche, KI für terroristische Aktivitäten zu nutzen. Im November 2023 berichtete Tech Against Terrorism, dass im Rahmen ihres Monitorings mehr als 5.000 KI-generierte Inhalte archiviert worden seien, die von Terrorist*innen und gewalttätigen Extremist*innen erstellt oder verbreitet wurden. Die Organisation beobachtete solche Experimente sowohl im Umfeld des Islamischen Staats und von Al-Qaida-Anhänger*innen als auch in Neonazi-Kreisen. So veröffentlichte beispielsweise eine dem Islamischen Staat nahestehende Gruppe für technischen Support eine Anleitung, in der sie ihre Netzwerke über die sichere Nutzung eines KI-Inhaltsgenerators informierte. Ein Anhänger des Islamischen Staats gab an, ein KI-Transkriptionstool verwendet zu haben, um eine Botschaft der Führungsriege zu transkribieren und zu übersetzen, während ein al-Qaida-nahes Medium Plakate mit Bildern veröffentlichte, die offenbar mithilfe von generativer KI erstellt worden waren. Im rechtsextremen Spektrum konnten Forscher*innen einen Telegram-Kanal mit KI-generierten neonazistischen, rassistischen und antisemitischen Bildern sowie einen „Leitfaden zur memetischen Kriegsführung" ausmachen, der Propagandist*innen dazu auffordert, Bildgeneratoren zur Erstellung extremistischer Memes zu nutzen (Tech Against Terrorism 2023b).

Die Zahl von 5.000 KI-generierten Inhalten sorgte zunächst für Besorgnis. Diese Inhalte machten nach Informationen von Tech Against Terrorism allerdings nur einem kleinen Teil der extremistischen Inhalte aus, die im Rahmen des Monitorings identifiziert wurden. Die Organisation fand kaum Anhaltspunkte dafür, dass generative KI zu diesem Zeitpunkt systematisch eingesetzt wurde, und stufte deren Nutzung daher als in erster Linie experimentell ein. Diese Unterscheidung ist nach wie vor von Bedeutung: Zwar belegt die Existenz Tausender synthetischer Bilder, dass Extremist*innen mit dieser Technologie experimentieren, doch lässt sich daraus nicht ableiten, wie viele einzelne Akteur*innen solche Bilder erstellt haben, ob sie Teil koordinierter Kampagnen waren oder ob sie ein großes Publikum erreicht und beeinflusst haben. Gleichwohl erwiesen sich einige extremistische Kreise bei diesen frühen Experimenten als besonders aktiv, wobei vor allem rechtsextreme und akzelerationistische Gruppierungen dabei hervorstachen.

So dokumentierten Federico Borgonovo, Silvano Rizieri Lucini und Giulia Porrino im Februar 2024 den Einsatz von generativer KI in einem dezentralen akzelerationistischen Netzwerk, das auf Telegram aktiv war. Das Netzwerk war formal keiner Organisation zugeordnet, gleichwohl standen seine Kanäle in Verbindung mit extremistischen Milieus aus dem Umfeld von Gruppen und Netzwerken wie Ouest Casual, The American Futurist und ehemaligen Mitgliedern der Atomwaffen Division. Zwei Kanäle, die KI-generierte rechtsextreme Bilder verbreiteten, hatten zusammen nahezu 8.000 Bilder in Umlauf ge-

bracht. Die Zahl dieser Bilder vermittelt einen Eindruck, wie sehr auch diese experimentelle Nutzung zur Verbreitung solcher Inhalte beiträgt.

Dabei reproduzierten diese Inhalte größtenteils bekannte rechtsextreme Narrative und Ästhetiken, brachten aber keine neuen ideologischen Themen ein. Es fanden sich beispielsweise idealisierte Darstellungen von Nazi-Soldaten, Bilder von maskierten und bewaffneten weißen Männern, antisemitische Verschwörungstheorien, rassistische Aufrufe zur Gewalt sowie Verweise auf einen vermeintlichen „Großen Bevölkerungsaustausch" und „Weißen Genozid".

Zudem wurde generative KI vielfach dazu eingesetzt, um etablierte rechtsextreme Memes wie „Pepe, der Frosch" und „Moon Man" in explizit rassistischen oder neonazistischen Kontexten neu zu gestalten. Dabei diente generative KI weniger als Quelle neuer ideologischer Inhalte, sondern wurde vielmehr als Instrument zur raschen Reproduktion, Modifizierung und visuellen Ausgestaltung von Narrativen und Symbolen genutzt, die bereits in der extremistischen Kultur verankert waren. Ihre Bedeutung lag nicht in ideologischer Innovation, sondern in dem Potenzial, optisch anspruchsvoll erscheinende Inhalte in großer Zahl zu produzieren und zu verbreiten (Borgonovo et al. 2024). Generell verweisen die vorliegenden Erkenntnisse eher auf eine zunehmend experimentelle Nutzung von generativer KI hin, als dass damit bisher ein grundsätzlicher Wandel terroristischer Aktivitäten verbunden wäre.

Jailbreaking

Mit der dokumentierten Nutzung von generativer KI durch extremistische Akteur*innen verbindet sich auch die Frage nach den bestehenden Schutzmechanismen, die einen solchen Missbrauch verhindern könnten. Angesichts der zunehmenden Hinweise auf extremistische Experimente mit generativer KI investieren Technologieunternehmen verstärkt in Sicherheitsmaßnahmen, die darauf abzielen, einen möglichen Missbrauch dieser Systeme einzudämmen (OpenAI 2023, 2024; Anthropic 2025, 2026; Google 2023). Bevor sie zum Einsatz kommen, durchlaufen KI-Modelle umfassende Sicherheitstrainings und werden mit Schutzmechanismen ausgestattet, die gemeinhin als Guardrails bezeichnet werden. Diese sollen verhindern, dass Inhalte generiert werden können, die Terrorismus, gewalttätigen Extremismus oder andere Formen von Gewalt fördern (Bai et al. 2022). Im Zuge der zunehmenden Leistungsfähigkeit generativer KI-Modelle haben Entwickler*innen diese Schutzmaßnahmen kontinuierlich weiterentwickelt und ihre Widerstandsfähigkeit gegen Missbrauch verstärkt (Shang et al. 2025). Dabei sind die Entwicklungen in Bezug auf Modellsicherheit uneinheitlich. So verweisen Studien darauf, dass selbst immer leistungsfähigere Systeme anfällig bleiben für feindliche Angriffe und Versuche, ihre Sicherheitsvorkehrungen zu umgehen (Liu et al. 2023).

Zu den bekanntesten Methoden der Umgehung dieser Sicherheitsvorkehrungen gehört das Jailbreaking: die gezielte Manipulation eines KI-Systems, um dessen Sicherheitsbeschränkungen auszuhebeln und Zugriff auf Inhalte zu erlangen, die das System andernfalls nicht bereitstellen würde. Untersuchungen haben gezeigt, dass Nutzer*innen mit entsprechender Motivation Schwachstellen in den Schutzmechanismen des Modells ausnutzen können, um schädliche oder verbotene Inhalte zu generieren (Molas & Lopes 2024). Gewalttätige Extremist*innen und terroristische Organisationen könnten solche Schwachstellen nutzen, um generative KI für die Produktion von Propaganda, Radikalisierung, Anwerbung oder Planung und Vorbereitung von Gewalttaten zu nutzen. Entscheidend ist daher die Widerstandsfähigkeit von Sicherheitsmaßnahmen und Schutzvorkehrungen gerade auch in solchen Fällen, in denen Nutzer*innen aktiv versuchen, diese Maßnahme zu umgehen. Dies gilt umso mehr für extremistische Akteur*innen, deren Nutzung von KI gezielt und beständig auf das Aufspüren von Schwachstellen in diesen Systemen ausgerichtet ist.

Es dürfte allerdings unrealistisch sein, Jailbreaks durch die Forschung zur KI-Sicherheit vollständig zu verhindern (Wei et al. 2023; Zou et al. 2023; Peng et al. 2024). Aktuelle Ansätze zielen vielmehr darauf ab, den technischen, zeitlichen und rechnerischen Auf-

wand zu erhöhen, der für eine erfolgreiche Umgehung dieser Mechanismen erforderlich ist, um so die Hürden für groß angelegte Jailbreaks zu erhöhen (Boreiko et al. 2024; Andriushchenko et al. 2024; Nasr et al. 2026). So hat beispielsweise Anthropic eingeräumt, dass es vermutlich unmöglich ist, Jailbreaks vollständig zu verhindern. Das Unternehmen betont dagegen die Bedeutung von Maßnahmen, die die Geschwindigkeit von erfolgreichen Angriffen abbremsen und die damit verbundenen Kosten erhöhen, um so das Erkennen und die Abwehr dieser Angriffe zu erleichtern und deren weitergehende Wirkung zu begrenzen (Anthropic 2026). KI-Sicherheit lässt sich insofern am besten als ein fortlaufender Anpassungsprozess verstehen, in dem Entwickler*innen die Sicherheitsvorkehrungen als Reaktion auf neu auftretende Schwachstellen kontinuierlich verstärken, während betrügerische Nutzer*innen immer neue Techniken entwickeln, um diese zu umgehen.

Eine weitere Herausforderung stellt dabei das Fehlen eines einheitlichen Sicherheitsstandards im gesamten Ökosystem der generativen KI dar (NIST 2024a; NIST 2026). Jeder Anbieter legt selbst fest, welche Aktivitäten und Ergebnisse durch seine Modelle untersagt sind, wie die Systeme auf unbefugte Anfragen reagieren und in welchem Umfang sie sich gegen Versuche wehren, diese Einschränkungen zu umgehen (Anthropic 2024; OpenAI 2024). Die Resilienz dieser Modelle bezüglich feindlicher Eingaben variiert daher erheblich, auch wenn sie zum Schutz gegen ähnliche Angriffe konzipiert wurden.

Eine Bewertung der KI-Firma PRISM Eval aus dem Jahr 2025 macht diese Unterschiede deutlich (Peigné-Lefebvre et al. 2025). Das in Paris ansässige Unternehmen für KI-Sicherheit hat 41 hochmoderne Sprachmodelle, darunter Claude, ChatGPT, Gemini und Grok, auf automatisierte Jailbreak-Versuche getestet. Obwohl fast alle getesteten Modelle letztlich dazu gebracht werden konnten, schädliche Inhalte zu erstellen, unterschied sich der dazu erforderliche Aufwand um das bis zu 300-Fache. Einige Modelle widerstanden Hunderten von gegen die Sicherheitsmechanismen ausgerichteten Prompts, ehe sie unzulässige Ausgaben erzeugten, während andere bereits nach wenigen Versuchen dazu gebracht werden konnten. Besonders deutlich wurden die Unterschiede in Bezug auf chemische, biologische, radiologische, nukleare und explosive Gefährdungsszenarien (CBRNE): Das Modell von Anthropic benötigte Berichten zufolge etwa 700 Schritte, ehe es den Anfragen nachgab, während andere Modelle bereits nach ein oder zwei Schritten dazu gebracht werden konnten, vergleichbare Unterstützung zu leisten.

Diese Ergebnisse deuten darauf hin, dass die Sicherheit eines Modells sich nicht allein danach beurteilen lässt, ob ein System letztendlich umgangen werden kann. Die für eine erfolgreiche Umgehung erforderlichen Ressourcen stellen einen wichtigen Aspekt der Widerstandsfähigkeit des Modells dar, da eine Erhöhung des Zeitaufwands, der Rechenressourcen und des technischen Fachwissens, die zur Überwindung von Sicherheitsvorkehrungen erforderlich sind, die Durchführbarkeit einer systematischen Ausnutzung einschränken kann – selbst dort, wo eine vollständige Verhinderung nach wie vor unmöglich ist.

Zugleich führen erhebliche Unterschiede zwischen den einzelnen Modellen zu einer strukturellen Anfälligkeit: Nutzer*innen, die die KI für schädliche Zwecke anfragen, stoßen je nach System, auf das sie zugreifen, auf unterschiedlich starken Widerstand. Diese Abweichung ist insbesondere in Kontexten mit böswilligen oder extremistischen Nutzer*innen von Bedeutung, die über vergleichsweise freizügige Systeme leichteren Zugang zu verbotenen Inhalten erhalten. Bei der Widerstandsfähigkeit von Modellen geht es daher nicht allein um die Frage, ob sich Sicherheitsmaßnahmen umgehen lassen, sondern auch darum, wie konsequent und vorhersehbar diese Sicherheitsvorkehrungen den Zugriff auf schädliche Funktionen in verschiedenen Systemen einschränken.

Dieses Problem gewinnt angesichts von Forschungen zu KI-Risiken an Bedeutung, die mit dem Einsatz erwähnter CBRNE-Stoffe, mit besonderen cybertechnischen Fähigkeiten oder mit handlungsbezogenen Risiken von autonomer Machtausübung oder Manipulation verbunden sind. Die potenzielle Nutzung modernster KI durch terroristische Akteur*innen hat allerdings bisher in der Forschung nur rela-

tiv wenig Aufmerksamkeit erhalten (Frontier Model Forum 2025; Phuong et al. 2024; Anthropic 2025; Tech Against Terrorism 2025). Die daraus resultierende Lücke geht über das Fehlen standardisierter Ansätze zur Bewertung der Robustheit hinsichtlich Jailbreaking hinaus. Zudem stellt sich die Frage, ob bestehende Sicherheitsrahmenbedingungen ausreichend berücksichtigen, auf welche Weise terroristische Akteur*innen KI-Systeme für ideologische, operative oder propagandistische Zwecke ausnutzen könnten.

Die Unterschiede in der Robustheit der Modelle und die bisher nur eingeschränkte Berücksichtigung des Risikos eines Missbrauchs durch Terrorist*innen legt den Schluss nahe, dass aktuelle Sicherheitsbewertungen zwar Rückschlüsse darauf zulassen, ob Sicherheitsvorkehrungen umgangen werden können; Aussagen darüber, unter welchen Bedingungen und mit welchem Aufwand ein Jailbreaking möglich ist, sind dabei aber nur bedingt möglich. Für die Bewertung der Sicherheit von KI ist daher nicht allein die Existenz von Schutzmaßnahmen entscheidend, sondern deren Resilienz und Ausrichtung auf konkrete missbräuchliche Nutzungsformen, wie sie von terroristischen Akteur*innen angestrebt werden könnten.

Fazit

Generative KI wird zunehmend in die Aktivitäten terroristischer und gewalttätiger extremistischer Akteur*innen integriert, insbesondere für Propagandazwecke und die Erstellung von Inhalten. Ihre Auswirkungen auf die operative Ebene sind jedoch bislang nur begrenzt nachweisbar. Die in diesem Kapitel erörterten Fälle deuten darauf hin, dass KI derzeit eher dazu genutzt wird, bestehende extremistische Praktiken zu beschleunigen und auszuweiten, als diese grundlegend zu verändern. Zugleich verdeutlicht die zunehmende Verbreitung von Jailbreaking, dass die Schutzmaßnahmen, die eine missbräuchliche Nutzung verhindern sollen, gegenüber hartnäckigen Nutzer*innen nach wie vor anfällig sind. Dabei bestehen erhebliche Unterschiede in der Widerstandsfähigkeit der einzelnen Modelle.

Diese Entwicklungen machen eine Lücke in der aktuellen Forschung zum Thema KI-Sicherheit sichtbar. Während CBRNE-, Cyber- und agentische Risiken große Aufmerksamkeit erhalten, sind die spezifischen Wege, auf denen terroristische Akteur*innen neueste KI-Systeme ausnutzen könnten, vergleichsweise wenig untersucht. Angesichts der Tatsache, dass extremistische Akteur*innen weiterhin mit diesen Technologien experimentieren, wird es von entscheidender Bedeutung sein, die Möglichkeiten, Grenzen und Schwachstellen dieser Technologien zu verstehen, um die Rahmenbedingungen für deren Regulierung und die Sicherheit im KI-Bereich weiterzuentwickeln und damit den Missbrauch dieser Technologien durch Terrorist*innen und gewalttätige Extremist*innen wirksam zu bekämpfen.

Literaturverzeichnis:

Andriushchenko, M., Croce, F. und Flammarion, N. 2024. *Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks.* arXiv. https://arxiv.org/abs/2404.02151 (abgerufen am: 23.09.2026).

Anthropic 2024. *Claude's Constitution.* https://www.anthropic.com/constitution (abgerufen am 23.09.2026).

Anthropic 2025. *Strategic Warning for AI Risk: Progress and Insights from Our Frontier Red Team.* https://www.anthropic.com/news/strategic-warning-for-ai-risk-progress-and-insights-from-our-frontier-red-team (abgerufen am 23.09.2026).

Anthropic 2026. *Fable 5's Cyber Safeguards and Our Jailbreak Framework.* https://www.anthropic.com/news/fable-safeguards-jailbreak-framework (abgerufen am 23.09.2026).

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. & Kaplan, J. 2022. *Constitutional AI: Harmlessness from AI Feedback*. arXiv. https://arxiv.org/abs/2212.08073 (abgerufen am 23.09.2026).

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. & Yang, D. 2023. „Identifying and Mitigating the Security Risks of Generative AI." *Foundations and Trends in Privacy and Security* 6 (1), S. 1–52. doi: 10.1561/3300000041.

Borgonovo, F., Lucini, S.R. & Porrino, G. 2024. „Weapons of Mass Hate Dissemination: The Use of Artificial Intelligence by Right-Wing Extremists." *GNET Insight.* https://gnet-research.org/2024/02/23/weapons-of-mass-hate-dissemination-the-use-of-artificial-intelligence-by-right-wing-extremists/ (abgerufen am 23.09.2026).

Boreiko, V., Panfilov, A., Voracek, V., Hein, M. & Geiping, J. 2024. *A Realistic Threat Model for Large Language Model Jailbreaks.* arXiv. https://arxiv.org/abs/2410.16222 (abgerufen am 23.09.2026).

Council of Europe. 2025. *Report on the Emerging Patterns of Misuse of Technology by Terrorist Actors.* Europarat, Lenkungsausschuss zur Bekämpfung des Terrorismus (CDCT). https://rm.coe.int/report-on-the-emerging-patterns-of-misuse-of-technology-by-terrorist-a/4880281a94 (abgerufen am 23.09.2026).

Frontier Model Forum. 2025. *Frontier Mitigations.* Frontier Model Forum. https://www.frontiermodelforum.org/technical-reports/frontier-mitigations/ (abgerufen am 23.09.2026).

Google 2023. *Google's AI Red Team: The Ethical Hackers Making AI Safer.* https://blog.google/innovation-and-ai/technology/safety-security/googles-ai-red-team-the-ethical-hackers-making-ai-safer/ (abgerufen am 23.09.2026).

Janjeva, A., Harris, A., Mercer, S., Kasprzyk, A. & Gausen, A. 2023. *The Rapid Rise of Generative AI: Assessing Risks to Safety and Security*. Centre for Emerging Technology and Security, Alan Turing Institute. https://cetas.turing.ac.uk/publications/rapid-rise-generative-ai (abgerufen am 23.09.2026).

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 1 (2026)*. Cambridge Programme on AI Science & Policy, University of Cambridge. https://casp.ac/reports/ai-enabled-terrorism (abgerufen am 23.09.2026).

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 49 (2026)*. doi: 10.1080/1057610X.2023.2259195 (abgerufen am 23.09.2026).

Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., Wang, K. & Liu, Y. 2023. *Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study*. arXiv. https://arxiv.org/abs/2305.13860 (abgerufen am 23.09.2026).

Molas, B. & 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 (abgerufen am 23.09.2023).

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. & Tramèr, F. 2026. „The Attacker Moves Second: Stronger Adaptive Attacks Bypass Defenses Against LLM Jailbreaks and Prompt Injections", *35. USENIX Security Symposium*: 1.467–1.486. https://www.usenix.org/system/files/usenixsecurity26-nasr.pdf (abgerufen am: 23.09.2026).

National Institute of Standards and Technology (NIST). 2024a. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.* https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence (abgerufen am 23.09.2026).

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.* https://www.nist.gov/news-events/news/2024/05/us-secretary-commerce-gina-raimondo-releases-strategic-vision-ai-safety (abgerufen am 23.09.2026).

OpenAI. 2023. *OpenAI's Approach to Frontier Risk.* https://openai.com/global-affairs/our-approach-to-frontier-risk/ (abgerufen am 23.09.2026).

OpenAI. 2024. *Improving Model Safety Behavior with Rule-Based Rewards.* https://openai.com/index/improving-model-safety-behavior-with-rule-based-rewards/ (abgerufen am 23.09.2026).

Peigné-Lefebvre, P., Feuillade-Montixi, Q., David, T. & Miailhe, N. 2025. *LLM Robustness Leaderboard v1: Technical Report.* https://arxiv.org/abs/2508.06296 (abgerufen am 23.09.2026).

Peng, A., Michael, J., Sleight, H., Perez, E. & Sharma, M. 2024. *Rapid Response: Mitigating LLM Jailbreaks with a Few Examples.* https://arxiv.org/abs/2411.07494 (abgerufen am 23.09.2026).

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. & Shevlane, T. 2024. *Evaluating Frontier Models for Dangerous Capabilities*. https://arxiv.org/abs/2403.13793 (abgerufen am 23.09.2026).

Shang, Z. & Wei, W. 2025. *Evolving Security in LLMs: A Study of Jailbreak Attacks and Defenses.* https://arxiv.org/abs/2504.02080 (abgerufen am 23.09.2026).

Tech Against Terrorism. 2023a. *Early Terrorist Adoption of Generative AI. Tech Against Terrorism.* https://techagainstterrorism.org/hubfs/Tech%20Against%20Terrorism%20Briefing%20-%20Early%20terrorist%20experimentation%20with%20generative%20artificial%20intelligence%20services.pdf (abgerufen am 23.09.2026).

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.* https://techagainstterrorism.org/gen-ai (abgerufen am 23.09.2026).

Wei, A., Haghtalab, N. & Steinhardt, J. 2023. *Jailbroken: How Does LLM Safety Training Fail?* https://arxiv.org/abs/2307.02483 (abgerufen am 23.09.2026).

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J.Z. & Fredrikson, M. 2023. *Universal and Transferable Adversarial Attacks on Aligned Language Models.* https://arxiv.org/abs/2307.15043 (abgerufen am 23.09.2026).

Autor*innenprofile

Dr. Friedhelm Hartwig

Dr. Friedhelm Hartwig arbeitet als wissenschaftlicher Mitarbeiter bei modus|zad im Handlungsfeld Islamismus und realisiert seit acht Jahren Monitoring-Projekte, deren Ergebnisse Fachkräfte der Prävention und politischen Bildung über aktuelle extremistische Aktivitäten auf Social Media informieren. Er studierte Islamwissenschaft, Arabistik und evangelische Theologie und forschte am Leibniz Zentrum Moderner Orient Berlin.

Kaiya Reisch

Kaiya Reisch arbeitet bei der BAG RelEx und beschäftigt sich dort inhaltlich vor allem damit, wie generative KI für extremistische Ansprache in sozialen Medien genutzt wird und welche Möglichkeiten sie für die Prävention bietet. Kaiya Reisch ist Kulturanthropologin (M.A., Universität Bonn und Universität Tel Aviv). Ihr Schwerpunkt liegt auf kollektiven Identitätskonstruktionen in gesellschaftlichen Diskursen und Konflikten sowie deren Wandel, unter anderem im Kontext von Radikalisierungsprozessen.

Heron Lopes

Heron Lopes ist wissenschaftlicher Mitarbeiter bei Stop Hate Brazil und OSINT-Analyst in einem privaten Unternehmen. Seine Forschung konzentriert sich auf politische Gewalt, Aufstände, und den Missbrauch von Technologien für extremistische Zwecke. Lopes hat mit führenden Forschungsinstitutionen zusammengearbeitet, u.a. mit dem International Centre for Counter-Terrorism (ICCT) in den Niederlanden, dem Uppsala Conflict Data Programme (UCDP) in Schweden, und Leiden University in den Niederlanden.

Impressum

KN:IX connect
Verbund Islamismusprävention und Demokratieförderung
Email: info@kn-ix.de
Web: www.kn-ix.de

Herausgeber:

modus|zad
Alt-Reinickendorf 25
13407 Berlin

Ansprechperson:

Elena Jung
elena.jung@modus-zad.de

Redaktion:

Götz Nordbruch, Elena Jung

KN:IX connect | Verbund Islamismusprävention und Demokratieförderung wurde von IFAK e. V., ufuq.de, modus|zad und der BAG RelEx ins Leben gerufen, um die Vernetzung, den Erfahrungstransfer und die fachliche Weiterentwicklung im Themenfeld Islamismusprävention voranzutreiben.

KN:IX connect hat sich zum Ziel gesetzt, die Islamismusprävention auf Bundesebene nachhaltig zu stärken und die Vernetzung von Akteur*innen im Themenfeld Islamismusprävention zu fördern. Darüber hinaus bündeln wir Kompetenzen und entwickeln die fachliche Expertise weiter, um Wissen und Methoden einem breiten Kreis zugänglich zu machen.

Mit KN:IX connect richten wir uns u. a. an Fachkräfte der Islamismusprävention sowie in angrenzenden Arbeitsfeldern sowie an Multiplikator*innen und Fachkräfte aus Schule, Wissenschaft, Verwaltung und (Sicherheits)Behörden sowie Landesdemokratiezentren und Partnerschaften für Demokratie. Darüber hinaus stehen wir Politik und Medien als Ansprechpartner*innen zur Verfügung.

KN:IX connect ist ein Verbund im Programmbereich Entwicklung einer bundeszentralen Infrastruktur im Bundesprogramm „Demokratie Leben!"

Trendreports sind das offizielle Publikationsformat von KN:IX connect. Sie dienen dazu, Trends und Herausforderungen im Feld der Islamismusprävention frühzeitig zu erkennen und systematisch aufzuarbeiten, um die gewonnenen Erkenntnisse für Akteur*innen zugänglich zu machen und den Austausch zwischen Forschung und Praxis zu fördern.

Für inhaltliche Aussagen und Meinungsäußerungen tragen die Publizierenden dieser Veröffentlichung die Verantwortung.

ISSN 3053-626X

Ein Verbund von

BAG RelEx | modus | zad | ufuq.de | IFAK e.V. – Verein für multikulturelle Kinder- und Jugendhilfe - Migrationsarbeit

Gefördert vom Bundesministerium für Bildung, Familie, Senioren, Frauen und Jugend

im Rahmen des Bundesprogramms Demokratie leben!