

Zwischen Innovation und Manipulation

Governance von KI-generierten Inhalten im digitalen Informationsraum

Generative KI senkt Aufwand und Kosten für die gezielte Produktion und Verbreitung von Desinformation, ermöglicht die Personalisierung von Propaganda und erschwert die Unterscheidung zwischen authentischen und synthetischen Inhalten. Das hat weitreichende Folgen für Demokratie und gesellschaftlichen Zusammenhalt. Extremistische Akteure setzen diese Technologien bereits strategisch für Rekrutierung, Propaganda und Zielgruppenansprache ein, etwa durch KI-generierte Avatare, Deepfakes und automatisierte Chatbots. Zugleich eröffnet KI auch Chancen für die Präventionsarbeit und kann verantwortungsvoll eingesetzt zur Stärkung von Demokratie beitragen. Mit dem DSA und dem AI Act wurden zwar erste regulatorische Grundlagen geschaffen, doch Durchsetzung, technische Standards und phänomenübergreifende Prävention bleiben vorerst unzureichend. Dieser Policy-Brief leitet daraus vier priorisierte Handlungsempfehlungen an die Bundespolitik ab: verbindliche Kennzeichnungs- und Herkunftsstandards, eine gestärkte Plattformaufsicht, ein phänomenübergreifend angelegtes Bundesförderprogramm für KI in der Extremismusprävention sowie eine Stärkung digitaler Medienkompetenz.

Jamuna Oehlmann
Geschäftsführerin
BAG ReEx



Kaiya Reisch
Referentin
BAG ReEx



Gefahren generativer KI für Demokratie und gesellschaftlichen Zusammenhalt

Künstliche Intelligenz (KI) hat innerhalb weniger Jahre die digitale Öffentlichkeit und demokratische Prozesse grundlegend verändert. Generative KI¹ ermöglicht die schnelle, kostengünstige und zielgruppenspezifische Produktion manipulativer Inhalte und verschiebt damit die Bedingungen demokratischer Kommunikation. Dadurch wird die Verifikation von Inhalten schwieriger. Besonders problematisch ist die Kombination aus KI-generierten Inhalten, datengetriebener Zielgruppenansprache und algorithmischer Verbreitung, weil sie die gezielte Beeinflussung großer und zugleich spezifischer Publika erleichtert und dadurch die Voraussetzungen informierter öffentlicher Meinungsbildung beeinträchtigen kann. Extremistische Akteure, darunter insbesondere auch islamistische, nutzen diese Technologien bereits strategisch zur Rekrutierung, Desinformation und Beeinflussung öffentlicher Debatten.² Obwohl erste regulatorische Grundlagen in Europa durch den Digital Services Act und den AI Act³ geschaffen wurden, bestehen weiterhin erhebliche Defizite bei Transparenz und Durchsetzung dieser Regelwerke sowie bei der technischen Nachvollziehbarkeit synthetischer Inhalte.

Die Debatte um KI und demokratische Öffentlichkeit ist damit längst keine Zukunftsfrage mehr. So warnte der Verfassungsschutz Baden-Württemberg kurz vor der dortigen Landtagswahl 2026 explizit vor dem Einsatz von Deepfakes und KI-gesteuerten Bots im Wahlkampf und verwies dabei auf konkrete Vorbilder aus dem Ausland.⁴ Wissenschaftler*innen warnen zudem vor einer neuen Generation sogenannter KI-Schwärme, also koordinierter Botnetzwerke, die insbesondere in sozialen Netzwerken und digitalen Debattenräumen mit variierenden Formulierungen auftreten können, um den Eindruck gesellschaftlicher Mehrheiten oder organischer Zustimmung zu erzeugen.⁵ Die technische Qualität solcher Inhalte erschwert ihre Erkennung zunehmend – sowohl für Nutzende als auch für Moderationssysteme und staatliche Institutionen. Generative KI senkt die Zugangshürden für digitale Einflussoperationen drastisch: Inhalte, deren Produktion früher

erhebliche technische Fähigkeiten und Ressourcen erforderte, lassen sich heute innerhalb weniger Sekunden erzeugen.

Besonders problematisch ist dabei die Kombination mehrerer technologischer Entwicklungen: generative KI zur Inhaltserstellung, datengetriebene Analyse- und Empfehlungssysteme, algorithmische Verstärkungsmechanismen sozialer Plattformen und automatisierte oder halbautomatisierte Botnetzwerke.⁶ Diese Elemente greifen ineinander und erzeugen neue Formen synthetischer Öffentlichkeit: Inhalte erscheinen authentisch, automatisierte Zustimmung oder Gegenrede wirkt organisch und Reichweite entsteht durch algorithmische Verstärkung.

Hinzu kommt die zunehmende Personalisierung manipulativer Kommunikation. Mithilfe datenbasierter Analysen können Narrative präzise auf bestimmte Zielgruppen zugeschnitten werden, etwa auf Jugendliche, politisch Entfremdete oder emotional vulnerable Nutzende. Soziale Plattformen wie TikTok begünstigen diese Entwicklung besonders, da ihre Empfehlungssysteme stark auf emotionale Interaktion optimiert sind.⁷

Damit verändert sich auch die Herausforderung demokratischer Governance grundlegend. Die Frage lautet nicht mehr allein, wie einzelne Falschinformationen entfernt werden können, sondern wie demokratische Öffentlichkeiten unter Bedingungen synthetischer Kommunikation geschützt werden können.

Die Erosion gemeinsamer Wirklichkeit

Die zentrale demokratische Herausforderung KI-generierter Inhalte liegt also nicht allein in einzelnen Falschinformationen, sondern in der schrittweisen Erosion gemeinsamer Wirklichkeitsgrundlagen durch die zunehmende Verbreitung synthetischer und schwer überprüfbarer Inhalte. Deepfakes und synthetische Medien verwischen die Grenze zwischen authentischen und manipulierten Inhalten. Gleichzeitig erzeugen koordinierte Botnetzwerke und künstliche Identitäten den Eindruck gesellschaftlicher Mehrheiten, die real nicht existieren.

Hinzu kommt die sogenannte Lügner-Dividende: Je glaubwürdiger KI-generierte Inhalte werden, desto leichter lassen sich auch echte Informationen als manipuliert diskreditieren. Dies untergräbt langfristig das Vertrauen in journalistische Institutionen, staatliche Kommunikation und digitale Informationsräume insgesamt.⁸

Algorithmische Plattformlogiken verstärken diesen Effekt zusätzlich. Plattformen priorisieren Inhalte nicht nach ihrem Wahrheitsgehalt, sondern nach ihrer Interaktionswahrscheinlichkeit. Polarisierende, emotionalisierende oder empörende Inhalte erzielen dadurch systematisch höhere Reichweiten.⁹ Die Folge ist eine Fragmentierung digitaler Öffentlichkeit in voneinander abgeschottete Teilöffentlichkeiten. Extremistische Narrative erscheinen sichtbarer und gesellschaftlich relevanter, als sie es tatsächlich sind. Gleichzeitig können sich Nutzende aus als aggressiv wahrgenommenen Debattenräumen zurückziehen, wodurch sich das Diskussionsklima weiter polarisiert. Sichtbarkeit orientiert sich zunehmend an emotionaler Reaktion statt an epistemischer Qualität.

Die besondere Gefahr durch islamistische Akteure

Dieses Kapitel zeigt am Beispiel islamistischer Akteure, wie generative KI strategisch für Rekrutierung, Propaganda und die gezielte Ansprache bestimmter Zielgruppen eingesetzt wird. KI wird zwar von Extremist*innen aller Richtungen genutzt, islamistische Gruppierungen wie der sogenannte Islamische Staat (IS) gelten jedoch als Early Adopters. Schon früh nutzten sie automatisierte Sprachübersetzung, um Inhalte in unterschiedliche Sprachen zu übertragen und so transnationale digitale Kommunikationsräume zu erschließen. Mittlerweile kommen KI-generierte Avatare, automatisierte Chatbots, synthetische Profilbilder und Deepfake-Inhalte hinzu. Dadurch lassen sich propagandistische Inhalte schneller produzieren, stärker emotionalisieren und zielgruppenspezifisch verbreiten, insbesondere an junge Nutzende auf visuell geprägten Plattformen wie TikTok.¹⁰

In der Praxis zeigt sich das bereits deutlich: Der IS setzte KI-generierte Avatar-Sprecher in Propagan-

davideos ein, während al-Qaida-nahe Netzwerke Anleitungen zur Nutzung von ChatGPT für propagandistische Zwecke veröffentlichten. KI-generierte Profilbilder, Deepfake-Videos oder automatisierte Kommunikationssysteme ermöglichen es so extremistischen Gruppen, professionelle Medienkampagnen mit vergleichsweise geringem Aufwand umzusetzen.¹¹ Der sogenannte Islamische Staat Provinz Khorasan (ISKAP) setzt darüber hinaus automatisierte Rekrutierungschatbots ein, die kontinuierlich über mehrere Plattformen hinweg mit potenziellen Rekrut*innen interagieren und ihre Ansprache an individuelle ideologische Neigungen und Vulnerabilitäten anpassen.¹² Solche privaten, geschlossenen Dialoge mit Chatbots und KI-Companions sind ein eigenständiges Radikalisierungsrisiko, das nicht an eine bestimmte Ideologie gebunden ist.¹³ Inzwischen wird der Einsatz von KI teilweise sogar fest in den Organisationsstrukturen verankert: Nachfolgeorganisationen von Boko Haram in Nigeria betreiben eigene „KI-Einheiten“, die über VPNs kommerzielle Modelle wie ChatGPT, Claude oder Gemini für ihre Propaganda- und Rekrutierungsarbeit verwenden.¹⁴

Zudem nutzen islamistische Akteure die algorithmische Logik der Plattformen gezielt aus. Sie setzen KI-Tools passend zur jeweiligen Zielgruppe und Plattform ein, um Anhänger*innen zu gewinnen und zu radikalisieren.¹⁵ Über ästhetisierte Kurzvideos, KI-generierte Bildwelten oder popkulturelle Referenzen vermitteln sie ideologische Inhalte niedrigschwellig und laden sie emotional auf. So kursierten etwa animierte Deepfake-Musikvideos, in denen bekannte Figuren wie SpongeBob Schwammkopf oder populäre Youtuber*innen islamistische Kampfgesänge singen. Solche Videos richten sich gezielt an eine junge Zielgruppe.¹⁶ KI verschärft diese Dynamik erheblich, weil sie die Produktion manipulativer Inhalte automatisiert, professionalisiert und in großem Maßstab ermöglicht. Was früher viel Personal erforderte, entsteht heute mit wenig Ressourcen, in großen Mengen und in zunehmend täuschend echter Qualität.

Diese Beispiele zeigen, wie konkret islamistische Akteure KI bereits einsetzen. Ein ausschließlich islamistisches Phänomen ist das jedoch nicht mehr. Vergleichbare Werkzeuge stehen ebenso

rechtsextremen und anderen extremistischen Akteuren zur Verfügung, die sie zunehmend auf ähnliche Weise nutzen. Wirksame Prävention setzt deshalb eine enge, phänomenübergreifende Zusammenarbeit voraus, damit Erkenntnisse zu KI-gestützten Taktiken ausgetauscht und Doppelstrukturen vermieden werden.

Chancen: KI für Prävention nutzen und Demokratie stärken

Trotz der genannten Herausforderungen eröffnet generative KI auch weitreichende Chancen für Demokratieförderung und Extremismusprävention. Ein reflektierter und verantwortungsvoller Einsatz von KI in der Präventionsarbeit kann helfen, den technologischen Vorsprung manipulativer Akteure zumindest teilweise auszugleichen. KI-gestützte Chat- und Beratungsangebote können niedrigschwellige Informations- und Unterstützungsangebote bereitstellen, während KI-gestützte Analyseverfahren große und heterogene Mengen digitaler Kommunikationsdaten („Big Data“) automatisiert auswerten. Dadurch lassen sich Muster wie Radikalisierungstendenzen, Desinformationskampagnen oder koordinierte Einflussnahmen erkennen, die mit klassischen manuellen Verfahren nur schwer identifizierbar wären. KI-gestützte Monitoring- und Frühwarnsysteme können Präventionsakteure dabei unterstützen, Entwicklungen in digitalen Räumen systematisch zu beobachten, Risiken besser einzuschätzen und evidenzbasierte Gegenstrategien zu entwickeln. Aktuelle Analysen aus der Präventionsarbeit zeigen zudem, dass KI nicht nur für Propaganda, die Simulation sozialer Nähe und die emotionalisierte Ansprache potenziell vulnerabler Nutzender eingesetzt werden kann,¹⁷ sondern auch skalierbares Prebunking¹⁸ die schnellere Erkennung extremistischer Inhalte und Kommunikationsmuster sowie Online-Streetwork durch Analyse-, Monitoring- und Dialogfunktionen unterstützen kann. KI ist damit nicht ausschließlich ein Instrument extremistischer Akteure, sondern kann Präventionsmaßnahmen auch ressourceneffizienter und besser skalierbar machen.

Dennoch bleibt für die Präventionsarbeit eine erhebliche strukturelle Asymmetrie bestehen: Während extremistische Inhalte automatisiert und

massenhaft produziert sowie zunehmend personalisiert verbreitet werden können, erfordern der Aufbau von Beziehungen, die Begleitung von Distanzierungsprozessen und die individuelle Ansprache vulnerabler Personen weiterhin erhebliche personelle und zeitliche Ressourcen. KI kann diese Arbeit unterstützen, sie jedoch nicht ersetzen.¹⁹

Bestehende Regulierungsansätze: Fortschritte und Grenzen

Mit dem Digital Services Act (DSA), der zwischen 2022 und 2024 schrittweise eingeführt wurde, sowie dem 2024 in Kraft getretenen und seither schrittweise anwendbaren AI Act hat die Europäische Union international vergleichsweise früh auf die Herausforderungen digitaler Plattformen und KI reagiert.

Der DSA verpflichtet insbesondere Very Large Online Platforms (VLOPs) zu Transparenzmaßnahmen, Risikobewertungen und systemischen Gegenmaßnahmen gegen Desinformation und Manipulation. Plattformen müssen unter anderem offenlegen, wie Empfehlungssysteme funktionieren und welche Maßnahmen gegen illegale oder manipulative Inhalte ergriffen werden.²⁰ Die Aufsicht über diese sehr großen Plattformen liegt maßgeblich bei der Europäischen Kommission; für die übrigen Anbieter sind nationale Koordinatoren für digitale Dienste zuständig, in Deutschland die Bundesnetzagentur. Der AI Act verfolgt einen risikobasierten Ansatz. Besonders risikoreiche KI-Systeme unterliegen strengeren Transparenz- und Kontrollpflichten. Seit dem 2. August 2026 gelten zudem die Transparenzpflichten nach Art. 50. Anbieter müssen KI-generierte Inhalte maschinenlesbar kennzeichnen, und Deepfakes müssen als solche offengelegt werden.²¹

Diese Regulierungen stellen wichtige Fortschritte dar, reichen jedoch bislang nicht aus. Erstens bestehen erhebliche Durchsetzungsprobleme: Sowohl der Europäischen Kommission als auch den nationalen Aufsichtsbehörden fehlen häufig ausreichende personelle, technische und finanzielle Ressourcen, um globale Plattformunternehmen effektiv zu kontrollieren. Zweitens fehlen bislang einheitliche technische Standards

zur Herkunftskennzeichnung synthetischer Inhalte. Eine maschinenlesbare Kennzeichnung ist zwar vorgeschrieben, wird aber weder einheitlich noch interoperabel umgesetzt; Wasserzeichen und Metadaten lassen sich zudem häufig leicht entfernen. Drittens bleibt die Verantwortungsverteilung zwischen Plattformen, Entwickler*innen und Nutzenden vielfach unklar. Plattformunternehmen argumentieren regelmäßig mit Meinungsfreiheit und Innovationsschutz, während gleichzeitig ihre algorithmischen Systeme manipulative Inhalte strukturell begünstigen. Viertens entwickelt sich die Technologie schneller als regulatorische Prozesse. Zwischen technischer Innovation und politischer Regulierung entsteht dadurch ein dauerhaftes Ungleichgewicht, ein „Governance Lag“. Fünftens verlieren DSA und AI Act ihre Wirkung bei frei zirkulierenden, quelloffenen („open-weight“) KI-Modellen: Deren Sicherheitsmechanismen lassen sich nachträglich entfernen, ohne dass danach noch ein verantwortlicher Anbieter greifbar bliebe.²²

Aus Sicht der Extremismusprävention zeigt sich zudem, dass bestehende regulatorische Ansätze häufig zu stark auf nachträgliche Inhaltsmoderation ausgerichtet sind. Resilienzorientierte, präventive Strategien wie etwa digitale Bildung, Prebunking und Früherkennung müssen daher langfristig stärker priorisiert werden.

Handlungsempfehlungen

1. Kennzeichnung und Herkunftsnachweise verbindlich standardisieren und durchsetzen

Die Kennzeichnungspflicht des AI Act braucht verbindliche, interoperable technische Standards. Deutschland sollte sich auf EU-Ebene für solche Kennzeichnungs- und Herkunftsstandards (Content Provenance, z. B. C2PA) einsetzen und ihre Durchsetzung national konsequent sicherstellen. Freiwillige Selbstverpflichtungen großer KI-Anbieter: Ohne rechtliche Verbindlichkeit bleiben Umsetzung und Zeitpunkt im Ermessen der Anbieter. Notwendig sind plattformübergreifende Standards, verpflichtende Offenlegung, technische Prüfschnittstellen für Faktenchecker*innen und Sicherheitsbehörden sowie klare Sanktionen. Das betrifft auch islamistische Propagandaformate wie Avatar-Videos,

deren synthetischer Ursprung für junge Zielgruppen kaum erkennbar ist. Zwar werden extremistische Akteure ihre Inhalte nicht freiwillig kennzeichnen, und mit Open-Weight-Modellen erzeugte Inhalte entziehen sich der Pflicht häufig. Je verbreiteter verlässliche Herkunftsnachweise aber sind, desto eher fallen Inhalte ohne Nachweis auf und werden von Plattformen und Nutzenden kritischer geprüft.

2. Plattformaufsicht und DSA-Durchsetzung stärken

Die zentrale Schwachstelle liegt nicht im Fehlen rechtlicher Grundlagen, sondern in deren Durchsetzung. Deutschland sollte die Bundesnetzagentur als Koordinator für digitale Dienste personell und technisch stärken, unabhängige Auditstrukturen und Forschungszugänge zu Plattformdaten ausbauen und sich auf EU-Ebene dafür einsetzen, dass die Europäische Kommission ihre Aufsicht über sehr große Plattformen mit ausreichenden Kapazitäten und wirksamen Sanktionen ausübt. Plattformen sollten verpflichtet werden, im Rahmen ihrer Risikobewertungen nach dem DSA regelmäßig offenzulegen, wie ihre Empfehlungsalgorithmen zur Verbreitung islamistischer und anderer extremistischer Propaganda an junge Nutzende beitragen, und wirksame Gegenmaßnahmen nachzuweisen. Bei viralen Hochrisikoinhalten, etwa in Wahlkampfphasen, braucht es verpflichtende menschliche Prüfmechanismen.

3. Bundesförderprogramm „KI in der Extremismusprävention“ auflegen

Gezielte Bundesförderung ist der wirksamste Hebel, um die Ressourcenasymmetrie zwischen automatisiert produzierter Propaganda und personalintensiver Präventionsarbeit auszugleichen. Sie macht KI für die Prävention nutzbar, statt sie allein extremistischen Akteuren zu überlassen. Das Programm sollte phänomenübergreifend angelegt sein und zivilgesellschaftlichen Präventionsträgern aller Phänomenbereiche, Sicherheitsbehörden und weiteren P/CVE-Akteuren den Zugang zu KI-Werkzeugen ermöglichen, etwa zu übertragbaren Prebunking-Kampagnen, Detektionswerkzeugen zur Früherkennung von Radikalisierungsindikatoren und chatbotgestütztem

Online-Streetwork. Verbindliche Evaluationsstandards sollten methodische Grenzen, etwa unklare Wirksamkeit oder Risiken der Anthropomorphisierung von Chatbots, von Beginn an mitdenken.

4. Medienkompetenz phänomenübergreifend stärken

Regulierung allein reicht nicht, Demokratien brauchen resilientere Informationskulturen. Der Bund sollte außerschulische Bildungsangebote ausbauen, Erwachsenenbildung zu KI und zu Desinformation fördern und zivilgesellschaftliche Präventionsträger, einschließlich der religiös begründeten Extremismusprävention, langfristig darin unterstützen, eigene KI-Kompetenzen für ihre Bildungs- und Präventionsarbeit auf- und auszubauen. Prebunking-Angebote sollten in Abstimmung mit anderen Phänomenbereichen entwickelt werden. Der in diesem Policy-Brief vertiefte Fokus auf islamistische Ansprachestrategien liefert dafür Ausgangsmaterial, das methodisch von Beginn an auf andere extremistische Spektren übertragbar angelegt sein sollte, etwa popkulturelle Referenzen oder vermeintlich harmlose Deepfake-Formate für junge Zielgruppen.

Da insbesondere Jugendliche auf visuell geprägten Plattformen wie TikTok zur Zielgruppe KI-gestützter extremistischer Propaganda (einschließlich islamistischer Propaganda) zählen, sollte bereits in der Schule vermittelt werden, wie sich KI-generierte extremistische Inhalte erkennen lassen. Da schulische Medienbildung in der Zuständigkeit der Länder liegt, empfiehlt sich statt einer Bundesregelung eine gezielte Befassung der Kultusministerkonferenz (KMK). Ziel sollte es sein, islamismussensible und phänomenübergreifend anschlussfähige Deepfake- und KI-Kompetenz verbindlich in länderübergreifende Rahmenlehrpläne aufzunehmen.

Endnoten

- [1] Generative KI bezeichnet den Teilbereich der künstlichen Intelligenz, der darauf spezialisiert ist, neue Inhalte wie Bilder, Videos oder Texte zu erschaffen.
- [2] Lang, I. (2026). Künstliche Intelligenz und Islamismus(-prävention). Eine Einordnung zum Verhältnis technologischer Innovation und islamistischer Radikalisierung im digitalen Raum. Infodienst Radikalisierungsprävention. Bundeszentrale für politische Bildung. <https://www.bpb.de/themen/infodienst/572545/kuenstliche-intelligenz-und-islamismus-praevention/> abgerufen [05.06.2026]; Scheeres, A. & Tabti, S. (2026). Künstliche Intelligenz in islamischen Online-Communities: zwischen „Early Adopters“ und KI-Skepsis. Einblicke für die Präventionspraxis (TPX Briefing 01/26). Thomasius Stiftung für internationale Bildung und Wissenschaft / TPX-Institut. <https://tpx-institut.de/wp-content/uploads/2026/04/TPX-Briefing-01-26-%E2%80%93-KI-in-der-Präventionsarbeit.pdf> abgerufen [05.06.2026].
- [3] Der Digital Services Act (DSA, seit 2022 schrittweise eingeführt) reguliert die Pflichten von Onlineplattformen im Umgang mit illegalen und manipulativen Inhalten; der AI Act (seit 2024 geltend) regelt risikobasiert den Einsatz von KI-Systemen in der EU. Dazu ausführlicher im Abschnitt „Bestehende Regulierungsansätze: Fortschritte und Grenzen“.
- [4] Verfassungsschutz Baden-Württemberg (2026). Die Landtagswahl 2026: KI-generierte Inhalte und Deepfakes im Wahlkampf – Hintergründe und Handlungsempfehlungen. <https://www.verfassungsschutz-bw.de/Lde/Startseite/Meldungen+und+Archiv/KI-generierte+Inhalte+und+Deepfakes+im+Wahlkampf/> abgerufen [27.07.2026].
- [5] Denzel, T. (2026). Wie Deepfakes und Bots die Demokratie gefährden. Tagesschau. <https://www.tagesschau.de/inland/innenpolitik/bots-fakes-wahlkampf-100.html> abgerufen [05.06.2026].
- [6] Denzel, T. (2026).
- [7] Schroeder, D. T. et al. (2026). How malicious AI swarms can threaten democracy. Science, 391(6783), 354–357.
- [8] Schroeder, D. T. et al. (2026).
- [9] Dametto, D.; Hechler, S. & von Scheve, C. (2026). Emotionally aligned and structurally sorted: how opinion groups divide on climate and asylum. PNAS Nexus, 5(3), pgag051; Kukaj, A. (2020). Algorithmen, Extremismus und die Aufmerksamkeitsindustrie. Zeitschrift für Semiotik, 42(3–4), 87–91.
- [10] Lang, I. (2026).
- [11] Lang, I. (2026).
- [12] 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 [29.09.2026].
- [13] Dukić, S.; Vela, N. & Venkataramakrishnan, S. (2026). Radicalisation in Closed Loops: Risks and Intervention Opportunities with AI Chatbots and Companions. Institute for Strategic Dialogue (ISD). <https://www.isdglobal.org/publication/radicalisation-in-closed-loops-risks-and-intervention-opportunities-with-ai-chatbots-and-companions/> abgerufen [29.09.2026].
- [14] 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 [29.09.2026].
- [15] Scheeres, A. & Tabti, S. (2026).
- [16] Lang, I. (2026).
- [17] Scheeres, A. & Tabti, S. (2026); Kukaj, A. (2020).
- [18] Prebunking ist die frühzeitige Vermittlung von Wissen über typische Manipulations- und Täuschungsstrategien.
- [19] Champion, H. (2026). Artificial Intelligence, Disengagement, and Terrorism Prevention: Opportunities and Challenges. Global Network on Extremism and Technology (GNET). <https://gnet-research.org/2026/05/18/artificial-intelligence-disengagement-and-terrorism-prevention-opportunities-and-challenges/> abgerufen [15.07.2026].
- [20] European Commission, Directorate-General for Communications Networks, Content and Technology (2026, 23. März). The Digital Services Act. Shaping Europe's Digital Future. <https://digital-strategy.ec.europa.eu/en/policies/digital-services-act/> abgerufen [05.06.2026].
- [21] European Commission, Directorate-General for Communications Networks, Content and Technology (2026, 27. Januar). AI Act. Shaping Europe's Digital Future. <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai/> abgerufen [05.06.2026].
- [22] 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 [29.09.2026].

Bundesarbeitsgemeinschaft religiös begründeter Extremismus e. V.

Oranienstraße 58, 10969 Berlin

Tel.: 030 92126289

E-Mail: info@bag-relex.de

www.bag-relex.de

[@bag_relex](https://www.instagram.com/bag_relex)

Herausgeberin: BAG RelEx e. V.

Autor*innen: Jamuna Oehlmann, Kaiya Reisch

Layout: Sven Schulze

Fotos: ©Michel Buchmann für BAG RelEx

Redaktionsschluss: 05. Oktober 2026

BAG RelEx steht für Bundesarbeitsgemeinschaft religiös begründeter Extremismus e. V. Wir sind ein gemeinnütziger Verein, der im November 2016 mit dem Gedanken gegründet wurde, eine Plattform für Vernetzung, fachlichen Austausch, inhaltliche Weiterentwicklung und Interessenvertretung der zivilgesellschaftlichen Träger im Arbeitsfeld der Demokratieförderung und Prävention von religiös begründetem Extremismus zu schaffen. Diese Aspekte sind nach wie vor Ziele und Grundlagen unserer Arbeit. Unser inhaltlicher Schwerpunkt liegt aktuell auf der Prävention von islamistischem Extremismus.

Als Dachorganisation von rund 40 Mitgliedsorganisationen in ganz Deutschland stehen wir für eine Vielfalt an Ansätzen und Methoden. Diese spiegeln die langjährigen Erfahrungen im Arbeitsbereich wider.

Wir verstehen uns als Plattform für zivilgesellschaftliche Akteure sowie als Schnittstelle zwischen Zivilgesellschaft, Politik, Medien und Öffentlichkeit.

