NETZHELDEN 60+

1. Monitoringbericht
(September 2026)

Künstliche Intelligenz und ihre Funktion in den Sozialen Medien

Ein Projekt von

Violence Prevention Network

in Kooperation mit

IZRD
Interdisziplinäres Zentrum für Radikalisierungsprävention und Demokratieförderung e.V.

Gefördert durch die

bpb: Bundeszentrale für politische Bildung

Künstliche Intelligenz und ihre Funktion in den Sozialen Medien

Künstliche Intelligenz ist in den vergangenen Jahren in aller Munde. Seit der Veröffentlichung des Chatbots ChatGPT 2022 haben viele Menschen selbst Erfahrungen mit der Nutzung von KI gemacht. Die **Einsatzmöglichkeiten sind dabei sehr vielseitig** und werden auch von antidemokratischen und extremistischen Akteurinnen und Akteuren genutzt. Der vorliegende Bericht thematisiert die Fragen: Was ist KI? Wie wird sie in den Sozialen Medien von extremistischen Gruppen eingesetzt? Was sind zentrale Herausforderungen und Gefahren? Und wie kann man KI-Inhalte erkennen bzw. mit ihnen umgehen?

Formen von KI

Unter dem Begriff **„Künstliche Intelligenz" (KI)** werden verschiedene Dinge zusammengefasst oder verstanden. Die meisten Formen von KI, die Menschen heute nutzen, sind sogenannte generative Modelle; sie erstellen also verschiedene Arten von Inhalten. Manche Modelle (Large Language Models, LLMs) sind auf das **Verarbeiten und Erstellen** von Texten ausgelegt. Dazu werden sie mit großen Datensätzen an Text trainiert und versuchen dann jeweils ein passendes nächstes Wort vorherzusagen. Dadurch können schnell große Mengen an Texten erstellt werden, die sprachlich und inhaltlich überzeugend wirken können. Andere KI-Modelle ermöglichen die Erstellung von Bildern und Videos, die teils täuschend echt aussehen (für mehr Informationen siehe BpB „Was ist KI und welche Formen von KI gibt es?").

Formen von KI

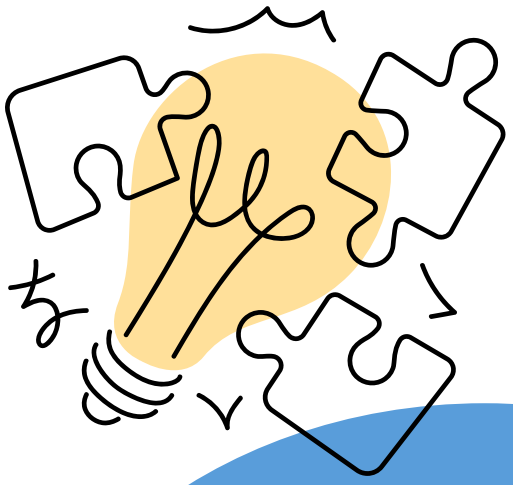
	Was kann sie?	Bekannte Beispiele
 **Text-KI**	Versteht und erstellt Texte, beantwortet Fragen, fasst Inhalte zusammen oder hilft beim Schreiben.	ChatGPT, Gemini, Claude, Microsoft Copilot
 **Bild-KI**	Erstellt Bilder aus Textbeschreibungen oder verändert vorhandene Bilder.	Midjourney, DALL·E, Gemini (Bildgenerator), Nano Banana 2
 **Video-KI**	Erstellt oder bearbeitet Videos anhand von Text oder Bildern.	Veo, Sora, Runway, Pika
 **Sprach-KI**	Erkennt gesprochene Sprache, wandelt Sprache in Text um oder liest Texte natürlich vor.	Whisper, ElevenLabs, Google Speech-to-Text
 **Musik-KI**	Komponiert Musik oder erzeugt Lieder und Instrumentalstücke.	Suno, Udio
 **Alltags-KI**	Unterstützt bei Übersetzungen, Suchanfragen, Navigation oder Produktempfehlungen.	DeepL, Google Translate & Maps, Amazon-Empfehlungen

In den Sozialen Medien gibt es in den letzten Monaten **immer mehr KI-generierte und teils relativ belanglose und inhaltsleere Beiträge** (auch als „KI-Slop“, engl. slop für „Abfall“, „Schlamm“ oder „Fraß“). Auch antidemokratische und extremistische Akteurinnen und Akteure nutzen die Möglichkeiten, die KI bietet, für ihre Zwecke. Einer der Anwendungsfälle, der sich am häufigsten beobachten lässt, ist die Erstellung von Bildern im gezeichneten oder animierten Stil. Diese sind häufig darauf ausgelegt, von anderen geteilt zu werden (sogenannte „Sharepics“ von engl. share = teilen und picture = Bild). Dabei werden bspw. rassistische Parolen, vermeintliche „Witze“ oder Bezugnahmen auf aktuelle Ereignisse mit einer Überschrift und einem **KI-generierten Bild** vermittelt.

Problematisch sind diese Bilder vor allem aufgrund ihrer inhaltlichen Botschaft, die bspw. migrantisch gelesene oder **queere** Menschen diskriminiert. Dank der KI-generierten Bilder und Videos können **Feindbilder einfach visuell dargestellt** werden.

Queer

Sammelbegriff für Menschen mit unterschiedlichen sexuellen Orientierungen und geschlechtlichen Identitäten, zum Beispiel lesbische, schwule, bisexuelle, trans- oder intergeschlechtliche Personen. Der Begriff wurde früher oft als Beleidigung verwendet, wird heute jedoch von vielen Menschen als positive Selbstbezeichnung genutzt.



Bei gezeichneten Bildern oder solchen, die sehr animiert aussehen, lässt sich meist gut erkennen, dass es sich nicht um ein echtes Foto handelt. Hier ist die **Verwechslungsgefahr** gering. Bei realistisch wirkenden Fotos gibt es schon länger das Problem, dass sich gefälschte Bilder (z. B. mit Fotobearbeitungsprogrammen) nur schwer auf den ersten Blick erkennen lassen. Teilweise sind Menschen bei Fotos daher bereits sensibilisiert, dass es sich auch um Fälschungen handeln kann.

Videos waren in der Vergangenheit ein besserer Indikator dafür, ob etwas echt ist, weil sie schwerer zu fälschen waren. KI-generierte **Videos, die täuschend echt aussehen**, sind deshalb besonders herausfordernd, da viele Personen sich auf visuelle Hinweise verlassen, um zu beurteilen, ob etwas echt ist.

KI und Verschwörungsmythen

KI-generierte Inhalte bieten gerade für Verschwörungsmythen die Möglichkeit, **Behauptungen, für die es keine Beweise gibt, mit visuellen Mitteln wie Bildern oder Videos vermeintlich zu belegen**. Wenn Menschen diese Videos auf Social Media sehen, überprüfen sie nicht immer, ob diese Videos echt sind und Erinnerungen an Videos und Bilder bleiben ggf. im Gedächtnis. Dies kann die Sichtbarkeit und Glaubwürdigkeit von Verschwörungsmythen erhöhen oder umgekehrt Personen verunsichern, was überhaupt echt ist.

Beispielhaft dafür steht ein kurzes Video auf *TikTok*, welches **Falschinformationen** verbreitet. Das KI-generierte Video suggeriert, dass der verstorbene Sexualstraftäter Jeffrey Epstein angeblich in Tel-Aviv beim Friseur ist und sein Aussehen verändert. Auf verschiedenen Social Media-Plattformen gab es (antisemitische) **Verschwörungs-mythen,** dass Jeffrey Epstein gar nicht tot sei. Das Video liefert einen vermeintlichen visuellen Beweis für eine derartige falsche Behauptung.

Ob das Video tatsächlich darauf abzielt, Personen zu täuschen, oder ob es den Erstellerinnen oder Erstellern vor allem darum geht, Aufmerksam-keit und damit Geld zu generieren, lässt sich nicht abschließend sagen.

Ein anderes Video, das auf *Facebook* über 73.000-mal angesehen wurde und rund 7.600 Likes erhielt, zeigt wie **KI-generierte Inhalte bewusst zur Täuschung eingesetzt werden**. Darin ist eine vermeintliche Demonstration einer großen Menschenmenge von AfD-Sympathisantinnen und Sympathisanten zu sehen, die „Aus Liebe zu Deutschland, wählen wir AfD!“ skandieren. Der Beschreibungstext behauptet, es hätten sich Tausende Menschen zu einer Demonstration in Berlin versammelt. Bei genauerem Hinsehen wird jedoch klar, dass das Video mit KI erstellt ist, die Gesichter der Menschen weisen Verzerrungen auf und die Personen heben unnatürlich synchron die Arme.

Auch der **Beschreibungstext liefert Hinweise dafür, KI-generiert zu sein**. Zu diesen Anzeichen zählen zum Beispiel sehr allgemeine Formulierungen, ein sehr abstrakter Text, der konkrete Fakten vermeidet, sowie Füll- und Übergangssätze, mit denen Atmosphäre erzeugt wird, die aber wenig neue Informationen liefern.

Der Account selbst verweist im obersten Kommentar auf einen Artikel einer **als Nachrichtenportal auftretenden Website**, der ebenfalls über die vermeintliche Großdemonstration berichtet und starke Hinweise darauf liefert, KI-generiert zu sein und keine echten Informationen wiederzugeben.

Im Kommentarbereich finden sich nur sehr wenige Kommentare, die kritisch darauf verweisen, dass es sich um mit KI-erstellte Videos mit falschen Informationen handelt. Viele Kommentare nehmen positiv Bezug auf das Video, scheinen es für echt zu halten und unterstützen dessen Botschaft.[1] Hier wird KI also eingesetzt, um täuschend (echt) falsche Informationen zu verbreiten und sie aufgrund der Bilder glaubwürdiger wirken zu lassen. Dieses Beispiel steht exemplarisch für die vielfältigen Versuche, Falschinformationen mit KI noch glaubwürdiger zu verbreiten. Diese Dynamik lässt sich auch bei anderen Themen und Ereignissen beobachten wie bspw. dem russischen Angriffskrieg auf die Ukraine, dem „Nahostkonflikt“, dem Krieg zwischen den USA und dem Iran oder etwa dem Thema Migration.

Wie auch Staaten versuchen, über Social Media Einfluss zu nehmen und bspw. Wahlen zu manipulieren, untersuchen wir im nächsten Monitoringbericht.

[1] Ob es sich dabei teils auch um Bot-Accounts, also keine echten Nutzerinnen und Nutzer, handelt, lässt sich nicht abschließend prüfen.

Probleme durch KI

Ein zentrales Problem von KI-Chatbots ist, dass diese teilweise **falsche Antworten** liefern (sogenannte „Halluzinationen"). Dadurch, dass andere Antworten tatsächlich stimmen, die Antworten leicht zu erhalten sind und plausibel wirken, werden sie teilweise nicht hinterfragt.
Im Rahmen von generativer KI, die sehr realistische Bilder und Videos (inkl. der Stimme einer Person) erstellt, kann es äußerst schwierig sein, auf den ersten Blick zu identifizieren, ob ein Bild oder ein Gesicht echt ist.

Test: KI-generiertes Bild

Das links gezeigte Porträt wurde nicht fotografiert, sondern testweise für diesen Report vollständig von einer KI erstellt. Grundlage war lediglich eine kurze Anweisung (auch „Prompt" genannt): „Erstelle das fotorealistische Porträtbild einer Person, die 70 Jahre alt ist, aus Deutschland kommt und sich gut im Internet auskennt."

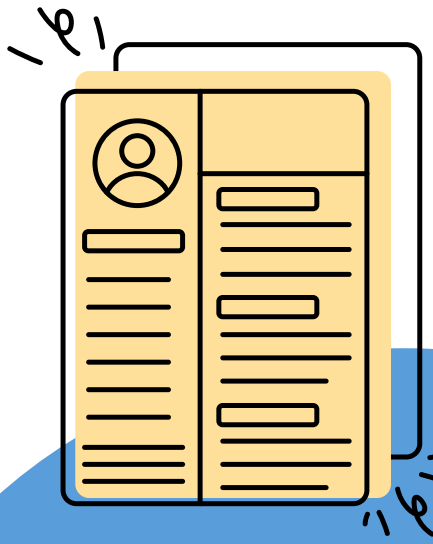
Das Ergebnis wirkt auf den ersten Blick wie ein echtes Foto. Schaut man jedoch genauer hin, lassen sich typische Fehler erkennen – beispielsweise unleserliche oder fehlerhafte Schrift auf dem Computerbildschirm und den Büchern.

Eine Untersuchung von Forscherinnen und Forschern des Massachusetts Institute of Technology (Link) liefert Hinweise darauf, dass viele Menschen nach dem Sehen von äußerst realistischen KI-Videos kurzzeitig in ihrer Wahrnehmung irritiert waren. Bei realen Videos, die im Anschluss an die KI-Videos gezeigt wurden, kam es bei den Personen teils zu Verunsicherung in ihrer Wahrnehmung und damit verbundenem Unwohlsein. Außerdem führte dies bei manchen Teilnehmenden dazu, dass sie sich den echten Menschen in anderen Videos kurzzeitig weniger verbunden fühlten. Dies verweist darauf, dass KI-generierte Inhalte auch beeinflussen, wie die Inhalte von echten Menschen wahrgenommen werden.

Im Kontext von Extremismus und Gewalttaten stellen sich zudem weitere Fragen. Beispielsweise, ob und wie KI-Chatbots Radikalisierungsprozesse beschleunigen können, weil sie z. B. die KI-nutzende Person immer weiter in ihrer Ansicht bestätigen*. Außerdem gibt es zunehmend Fälle, bei denen Attentäterinnen und Attentäter KI nutzten (Link), um Anschläge „besser“ zu planen oder Waffen bzw. Sprengsätze herzustellen.

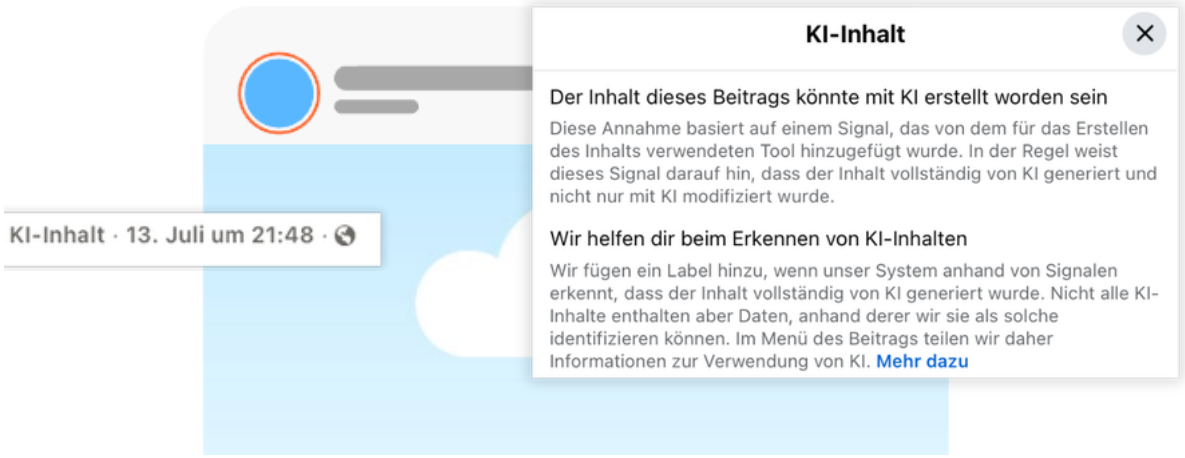
---

*Dabei handelt es sich um den sog. Sycophancy-Effekt, der die Tendenz eines KI-Systems bezeichnet, Nutzer*innen übermäßig zuzustimmen oder deren Aussagen zu bestätigen, statt sie kritisch zu hinterfragen oder gegebenenfalls zu widersprechen.



KI erkennen

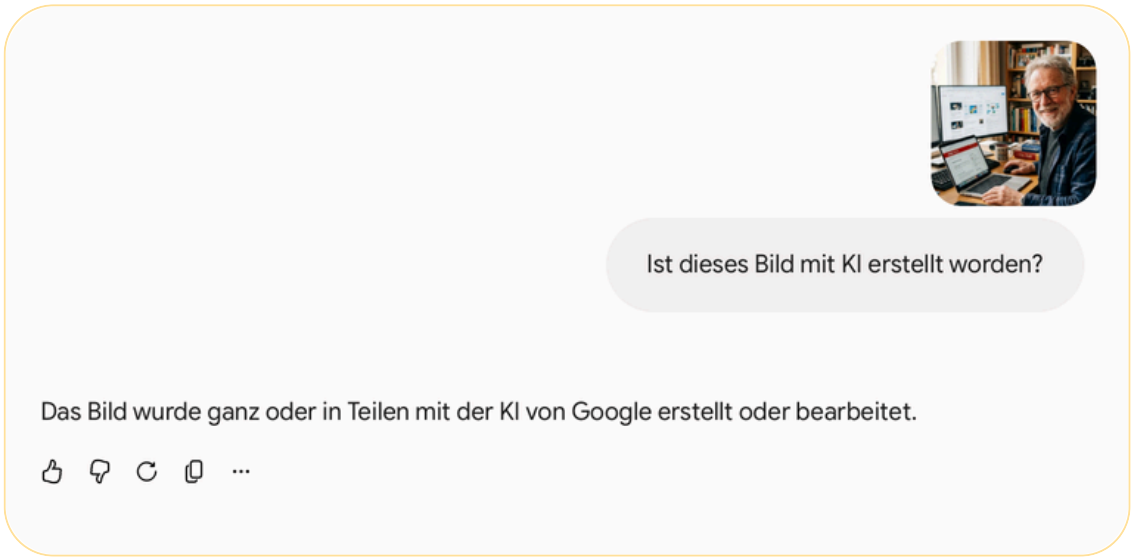
KI kann also ein Risiko darstellen. Daher stellt sich die Frage: Was kann man tun, um KI-Inhalte zu erkennen und wie geht man mit ihnen um? Viele KI-Inhalte haben keine Täuschungsabsicht und lassen sich sehr eindeutig als KI-generiert erkennen. Auf manchen Plattformen wie *Facebook* oder *Instagram* wird teilweise angezeigt, ob ein Video KI-generierte Inhalte enthält (dann erscheint dort der **Hinweis „KI-Info“**). Allerdings werden nicht alle KI-generierten Inhalte so erkannt bzw. gekennzeichnet.



Grundsätzlich hilft es auch, im Umgang mit KI-generierten Inhalten (ebenso wie mit nicht KI-generierten Inhalten) **zunächst kritisch deren Glaubwürdigkeit zu hinterfragen**. Außerdem kann auf visuelle oder sprachliche Hinweise geachtet werden, die **Anzeichen für KI** liefern: Beispielsweise Verzerrungen, unnatürliche Darstellungen von menschlichen Gesichtern, Händen oder Schrift, die keine echten Buchtstaben formt. Wenn Unsicherheit besteht, ob etwas wirklich echt ist, sollten **unabhängige Quellen** gesucht werden, um zu überprüfen, ob die Aussage stimmt oder ein Ereignis wirklich passiert ist.

KI erkennen mit KI

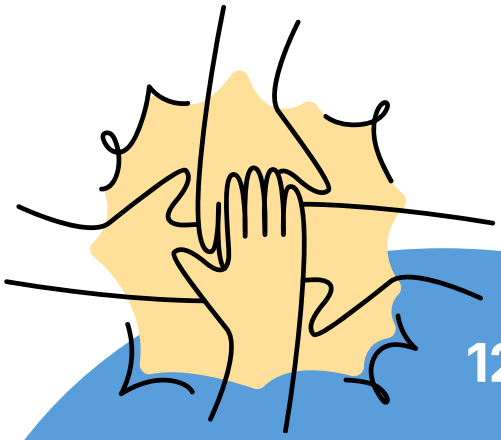
Bei dem KI-Assistenten von Google (Gemini) gibt es die Möglichkeit, ein Bild hochzuladen und zu fragen, ob dieses mit KI erstellt wurde. Wenn das Bild mit Gemini selbst erstellt wurde, kann das Programm ein digitales Wasserzeichen auslesen und dies eindeutig feststellen.



Ansprache KI-Fälschung

Wenn man in Diskussionen in Kommentarbereichen feststellt, dass Personen eine KI-Fälschung nicht erkannt haben, kann ein freundlicher, empathischer Hinweis darauf, dass etwas KI-generiert ist, hilfreich sein.

Allerdings gilt es, die **Personen nicht bloßzustellen**, **sondern konstruktiv zu helfen**, da es vielen Menschen schwerfällt, KI-generierte Inhalte zu identifizieren und dies kein Grund für Scham oder Ärger sein sollte.



Fazit

Abschließend lässt sich festhalten, dass KI-Inhalte auf Social Media immer häufiger werden und teils schwer zu erkennen sind. Auch antidemokratische und extremistische Akteurinnen und Akteure nutzen KI, um Falschinformationen zu verbreiten, Feindbilder darzustellen oder für sich selbst zu werben.
KI-Inhalte können Menschen verunsichern, wenn unklar scheint, ob etwas echt ist. Aber es gibt auch Möglichkeiten und Strategien, damit umzugehen und KI für positive Zwecke zu nutzen.

Weiterführende Materialien

Weitere Tipps und Ratgeber zur Identifikation von und zum Umgang mit KI gibt es hier:

- Artikel und Interaktives Quiz: Hinweise und Beispiele für die Erkennung KI-generierter Bilder (Link)
- Video: Deepfake Selbstexperiment: Wie easy kann ich Fake News mit KI erstellen? (Link)
- Artikel: Fake oder Wirklichkeit: Wieso und wie leicht lassen wir uns täuschen? (Link)
- Podcast: KI, Deepfakes & Co: Wie erkenne ich KI-Content? (Link)
- Quiz: Deepfake Detectives - Kannst Du den Deepfake erkennen? (Link)
- Faktencheck: Überprüfung von (Falsch-)Meldungen (Link)

Impressum

Herausgegeben von
NetzHelden 60+

Violence Prevention Network gGmbH
Alt-Reinickendorf 25
13407 Berlin
Tel.: (030) 917 05 464
post@violence-prevention-network.de
www.violence-prevention-network.de

Eingetragen beim
Amtsgericht Berlin-Charlottenburg
unter der Handelsregisternummer:
HRB 221974 B

https://violence-prevention-network.de/angebote/projektuebersicht/netzhelden-60plus/ und

https://www.izrd.de/de/netzhelden60plus.html

netzhelden@violence-prevention-network.de und

netzhelden@izrd.de

@netzhelden60plus

@netzhelden60plus

Ein Projekt von

Violence Prevention Network

in Kooperation mit

IZRD Interdisziplinäres Zentrum für Radikalisierungsprävention und Demokratieförderung e.V.

Gefördert durch die

bpb: Bundeszentrale für politische Bildung