ISD | Powering solutions to extremism, hate and disinformation

Generative KI und die deutsche extreme Rechte: Narrative, Taktiken und digitale Strategien

Von: Anna Hiller und Pablo Maristany de las Casas

ISD | Powering solutions to extremism, hate and disinformation

Amman | Berlin | London | Paris | Washington DC

www.isdglobal.org

Inhaltsverzeichnis

Zur Durchführung dieser Untersuchung beobachteten ISD-Analyst*innen Posts mit generativer künstlicher Intelligenz (KI) von offiziellen Konten der Alternative für Deutschland (AfD) und den Social-Media-Konten anderer rechtsextremer Akteure in Deutschland. Diese Untersuchung ergab, dass generative KI es diesen Akteur*innen ermöglicht, etablierte Narrative nahtlos in maßgeschneiderte Social-Media-Strategien zu integrieren. KI ist keine "Wunderwaffe" im Spielbuch der Rechtsextremen, sondern vielmehr eine leistungsstarke Ergänzung zu diesen etablierten Strategien. Dies ermöglicht es rechtsextremen Gruppen, große Mengen an ansprechendem Material auf kosten- und zeiteffiziente Weise zu erstellen und dabei die mangelnde Einhaltung EU-Gesetzes über digitale Dienste (DSA) durch Plattformen auszunutzen.

Wichtigste Erkenntnisse

- **Das ISD hat** seit April 2023 **883 Beiträge von deutschen rechtsextremen Konten identifiziert, die KI-generierte Inhalte (AIGC) enthalten**.

- **Die Partei Alternative für Deutschland (AfD), die bei der Bundestagswahl im Februar 2025 voraussichtlich zweitstärkste Kraft werden wird, wurde als eine der Hauptquellen für AIGC beobachtet.** Diese Inhalte stammen vom Hauptkonto der Partei, von AfD-Konten auf Landesebene und von einzelnen AfD-Politiker*innen auf Bundes- und Landesebene. AIGC-Inhalte wurden auf Facebook, Instagram, X, TikTok und YouTube gepostet. Im Oktober 2024, dem letzten untersuchten Monat, veröffentlichten AfD-Accounts auf allen Plattformen mehr als 50 Beiträge mit AIGC.

- **Rechtsextreme Community-Gruppen auf Facebook und rechtsextreme Musikkanäle auf YouTube machen ebenfalls regen Gebrauch von AIGC**. Die von ihnen erstellten Inhalte sowie das von der AfD produzierte Material werden in großem Umfang von Einzelpersonen auf Facebook, Instagram, X und TikTok geteilt und erneut gepostet, die keine offizielle Verbindung zu organisierten rechtsextremen Gruppen haben.

- **Zu den häufigen Narrativen, die im AIGC gefunden wurden, gehören neben Angriffen auf Geflüchtete, Einwander*innen, LGBTQ+- und Klima-Aktivist*innen und Oppositionsparteien zudem Inhalte, die Deutschland als starkes Land idealisieren, das bedroht ist und gerettet werden muss.** Inhalte, die zur "Remigration" aufrufen - die großflächige Abschiebung ethnischer Minderheiten unabhängig von ihrem Einwanderungsstatus - gehören zu den Narrativen, die die höchste Zustimmung erhielten.

- **KI-generierte Bilder, Memes und Lieder werden verwendet, um bei Mitgliedern rechtsextremer Gruppen und Anhängern rechtsextremer Seiten ein Gefühl der Identität zu schaffen.** In der Stichprobe von 883 Beiträgen fand das ISD 102 KI-generierte rechtsextreme Musikvideos.

- **Es wurde beobachtet, dass rechtsextreme Nutzende generative KI nutzen, um ihre Botschaften zu illustrieren und Videosequenzen und Bilder zu erstellen, um Szenen darzustellen, für die möglicherweise keine realen Bilder verfügbar sind, z. B. vermeintliche Kriminalität von Migrant*innen.**

- **Das ISD hat die Profile von drei weiblichen "Influencerinnen" identifiziert, die mit Hilfe generativer KI erstellt wurden.** Die Profile teilen Bilder und Reels mit rechtsextremen Inhalten und beziehen Stellung zu aktuellen Themen aus weiblicher Sicht, um eine parasoziale Beziehung zu ihrem Publikum aufzubauen, indem sie persönliche Informationen teilen und ein falsches Gefühl von Intimität erzeugen.

- **Rechtsextreme Akteur*innen nutzen die mangelnde Einhaltung der EU-Verordnung über digitale Dienste (DSA) seitens der Plattformen, um bestimmte AIGC zu kennzeichnen, und die begrenzten Möglichkeiten des KI-Gesetzes, die Verbreitung von AIGC zu stoppen.** In der Zeitspanne von einem Monat nach Meldung wurde keiner der AIGC-Beiträge in unserem Datensatz von den Plattformen gekennzeichnet. Nur 4 Prozent der Beiträge waren zum Zeitpunkt der Erstellung dieses Berichts als KI-generiert gekennzeichnet.

Einführung

Für diese Untersuchung sammelte das ISD insgesamt 883 Beiträge von 92 Konten mit AIGC, die von AfD-nahen rechtsextremen Akteur*innen auf Facebook, Instagram, X (früher Twitter), TikTok und YouTube zwischen April 2023 und November 2024 veröffentlicht wurden (siehe unten für die detaillierte Methodik). Frühere ISD-Forschungen hatten bereits die in Teilen als gesichert rechtsextrem eingestufte Partei Alternative für Deutschland (AfD) als produktiven Nutzende dieser Technologie identifiziert, was uns dazu veranlasste, die Nutzung generativer KI durch die AfD genauer zu untersuchen.

Die Nutzung von KI-generierten Inhalten (AIGC) durch die AfD wurde in den deutschen Medien erstmals im März 2023 erwähnt, als der Abgeordnete Norbert Kleinwächter ein KI-generiertes Bild postete, das Migranten als öffentliche Bedrohung darstellte, obwohl das erste KI-generierte Bild der AfD im August 2022 auf seiner Facebook-Seite veröffentlicht wurde. Seitdem haben rechtsextreme Akteur*innen generative KI eingesetzt, um Inhalte für Social-Media-Kampagnen zu erstellen. Offizielle AfD-Konten nutzen AIGC plattformübergreifend, um Bilder und kurze Videosequenzen für YouTube-Kurzfilme zu erstellen. Die Bilder und kurzen Videos stellen Einwander*innen als Bedrohung für "einheimische" Deutsche dar – diese typischerweise als blond und blauäugig dargestellt - und machen sich über die etablierten politischen Parteien lustig.

Eine Untersuchung der Süddeutschen Zeitung (SZ) hat ergeben, dass die AfD bei der Produktion von KI-Inhalten häufig mit der Medienagentur Tannwald Media zusammenarbeitet, die von Alexander Kleine gegründet wurde und ihm gehört, einem bekannten Akteur der deutschen Extremen Rechten. Die SZ-Untersuchung deckte Verbindungen zwischen Tannwald Media, der AfD und anderen Seiten auf, die rechtsextreme KI-Inhalte verbreiten, die im Folgenden erläutert werden.

Während die AfD im Mittelpunkt dieses Trends steht, setzen auch andere rechtsextreme Akteur*innen zunehmend auf AIGC. Dazu gehören Informationskanäle der "Neuen Rechten" wie die Junge Freiheit, rechtsextreme Online-Communities und individuelle Accounts. In

diesem Bericht wird untersucht, wie verschiedene Gruppen rechtsextremer Akteur*innen AIGC übernommen haben und welche Narrative und Taktiken sie verwenden.

In diesem Bericht sticht TikTok als eine Plattform hervor, auf der rechtsextremes AIGC erfolgreich verbreitet wird, was durch den Umfang des rechtsextremen AIGC und die einzigartige Verbreitungsstrategie der AfD auf der Plattform belegt wird. In einer Studie von 2024 bezeichnete die Bildungsstätte Anne Frank TikTok als ein "Paralleluniversum", das von den Rechtsextremen zur Verbreitung von Extremismus genutzt wird. In TikTok-Videos präsentiert sich die AfD als Retterin Deutschlands und vor allem der jungen Menschen. Forscher haben gemeinsame Taktiken wie "provokative Sprache, verschwörerische Rhetorik und Angstmacherei, die komplexe Themen vereinfachen und den rationalen Diskurs untergraben" festgestellt.

Obwohl das Konto der AfD-Partei auf Bundesebene im Mai 2022 von TikTok gesperrt wurde, teilen einzelne Politiker*innen und Fraktionen auf Landesebene weiterhin Inhalte. Die AfD hat auch von zahlreichen nicht angeschlossenen rechtsextremen Konten profitiert, die ihre Inhalte weiterverbreiten, von Konten, die andere rechtsextreme Inhalte mit dem Hashtag #afd teilen, und von Konten, die die Menschen dazu aufrufen, bei Landtags- und Bundestagswahlen für die AfD zu stimmen. AfD-Anhänger*innen und rechtsextreme Nutzer*innen verstärken die Wirkung von Inhalten, indem sie Beiträge liken, teilen und speichern und mit blauen Herzen (der Signaturfarbe der AfD) kommentieren. Rechtsextreme Inhalte werden auch häufig von anderen Konten herunter- und wieder hochgeladen, um ihre Reichweite zu erhöhen. AfD-Politiker*innen und andere Urheber*innen rechtsextremer Inhalte fordern ihre Anhänger*innen auf, diese Engagement-Strategie zu nutzen, um die Sichtbarkeit zu maximieren und sicherzustellen, dass die Inhalte auf der Plattform verfügbar bleiben, auch wenn das Konto, das sie ursprünglich geteilt hat, gelöscht wurde. Generative KI ermöglicht es der AfD und ihren Anhänger*innen, mehr rechtsextreme Inhalte zu produzieren und zu teilen, als dies ohne den Einsatz von KI möglich wäre.

In Anbetracht dieser Entwicklungen stellen sich Fragen zur Eignung und Interoperabilität von Vorschriften wie dem Gesetz über digitale Dienste (DSA), dem Gesetz über künstliche Intelligenz (AI), der Verordnung über terroristische Online-Inhalte (TCO) und den Leitlinien für Anbieter von VLOPs und VLOSEs zur Minderung systemischer Risiken bei Wahlen im Umgang mit der Verbreitung von AIGC in sozialen Medien. Die Hauptanliegen reichen von der Sicherstellung, dass Nutzende, die AIGC einsetzen, die Grundrechte, die Sicherheit und ethische Grundsätze respektieren, bis hin zur Bewertung der Angemessenheit für politische Zwecke sowohl on- als auch offline.

Nach einer tieferen Analyse der AIGC-Erkennung, der rechtsextremen Online-Akteur*innen, die sie einsetzen, und der Narrative, die sie fördern, zielt diese Untersuchung darauf ab, das Zusammenspiel der oben genannten Vorschriften zu bewerten. Folglich wird sie Empfehlungen vorschlagen, um etwaige Lücken zu schließen und die negativen Auswirkungen des rechtsextremen AIGC abzuschwächen.

Glossar

Desinformation
Falsche, irreführende oder manipulierte Inhalte, die in der Absicht verbreitet werden, zu täuschen oder zu schaden.

Diskriminierende Sprache
Sprache, die Personen aufgrund persönlicher Merkmale diskriminiert, was zu Marginalisierung und Ausgrenzung führen kann.

Fehlinformationen
Fehlinformationen sind falsche, irreführende oder manipulierte Inhalte, die unabhängig von der Absicht, zu täuschen oder zu schaden, verbreitet werden.

Gender Mainstreaming
Gender Mainstreaming hat sich international als Strategie zur Verwirklichung der Gleichstellung der Geschlechter durchgesetzt. Dabei geht es um die Einbeziehung einer geschlechtsspezifischen Perspektive in die Vorbereitung, Gestaltung, Umsetzung, Überwachung und Bewertung von Politiken, Regulierungsmaßnahmen und Ausgabenprogrammen, um die Gleichstellung von Frauen und Männern zu fördern und Diskriminierung zu bekämpfen.

Generative KI
Generative KI-Systeme basieren auf Deep-Learning-Modellen, die auf Rohdaten trainiert werden, zu denen Bücher, Artikel, Webseiten, Wikipedia-Einträge und aus dem Internet gesammelte Bilder gehören können. Diese Modelle sind darauf ausgelegt, statistische Muster in ihrem Trainingsdatensatz zu erkennen und "auf Aufforderung statistisch wahrscheinliche Ausgaben zu erzeugen", die den Daten, auf denen sie trainiert wurden, ähnlich, aber nicht identisch sind.[1] Dieser Bericht konzentriert sich auf Beispiele für generative KI-Systeme, die zur Erzeugung von synthetischen Texten, Bildern, Audio- und Videodaten verwendet werden können.

Gesetz über digitale Dienste (DSA)
Das DSA ist eine Verordnung der Europäischen Union, die Online-Vermittler und Plattformen wie Marktplätze, soziale Netzwerke, Content-Sharing-Plattformen, App-Stores, Suchmaschinen, Online-Reise- und Unterkunftsplattformen betrifft. Ihr Hauptziel ist es, illegale und schädliche Aktivitäten im Internet zu verhindern und die Verbreitung von Desinformationen zu stoppen. Sie gewährleistet die Sicherheit der Nutzenden, schützt die Grundrechte und schafft ein faires und offenes Umfeld für Online-Plattformen.

Gezielte Belästigung
Belästigung, die sich gegen eine bestimmte Person oder Gruppe richtet und die Absicht hat, diese zu bedrohen, zu provozieren oder in Bedrängnis zu bringen.

---

[1] IBM Forschung. (2023). Was ist generative KI? Abrufbar unter: https://research.ibm.com/blog/what-is-generative-AI.

Künstliche Intelligenz (KI)

Das ISD folgt der OECD-Definition von KI als "maschinengestütztes System, das für explizite oder implizite Ziele aus den Eingaben, die es erhält, ableitet, wie es Ergebnisse wie Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erzeugen kann, die physische oder virtuelle Umgebungen beeinflussen können. Verschiedene KI-Systeme unterscheiden sich in ihrem Grad an Autonomie und Anpassungsfähigkeit nach dem Einsatz".[2]

Neue Rechte (Neue Rechte)

Das Bundesamt für Verfassungsschutz (BfV) definiert die Neue Rechte als "ein informelles Netzwerk von Gruppen, Einzelpersonen und Organisationen, von nationalkonservativ bis rechtsextrem, die zusammenarbeiten, um ihre zum Teil antiliberalen und antidemokratischen Positionen in der Gesellschaft und im politischen Raum zu vertreten. Die parlamentarischen und außerparlamentarischen Bewegungen und die metapolitische Theorie und Praxis, mit denen das Netzwerk versucht, die vorpolitische Sphäre zu beeinflussen und die Grundlage für eine erfolgreiche politische Umsetzung ihrer antidemokratischen Positionen zu schaffen, sind eng mit dem Einsatz von Protesten und Demonstrationen verknüpft. Die Galionsfiguren der Neuen Rechten sind untereinander gut vernetzt und erfüllen innerhalb dieses Netzwerks unterschiedliche, teilweise komplementäre Rollen, um eine "Kulturrevolution von rechts" zu bewirken, indem sie unterschiedliche Zielgruppen ansprechen."

Rechtsextremismus

Das ISD definiert Rechtsextremismus als "eine Form des Nationalismus, die durch ihren Bezug auf rassische, ethnische oder kulturelle Vorherrschaft gekennzeichnet ist". In Übereinstimmung mit dem Wissenschaftler und Rechtsextremismusexperten Cas Mudde definiert das ISD Rechtsextremismus Gruppen und Einzelpersonen, die mindestens drei der folgenden Merkmale aufweisen: Nationalismus, Rassismus, Fremdenfeindlichkeit, Demokratiefeindlichkeit oder starkes Eintreten für den Staat.

Verschwörungserzählungen

Verschwörungserzählungen sind Versuche, Ereignisse oder Umstände als Ergebnis einer geheimen Verschwörung zu erklären, die von angeblichen (in der Regel elitären) Verschwörer*innen in böser Absicht inszeniert wurde.

[2]Russell, S., Perset, K., & Grobelnik, M. (November 29, 2023). Aktualisierungen der OECD-Definition eines KI-Systems erklärt. Organisation für wirtschaftliche Zusammenarbeit und Entwicklung. https://oecd.

Methodik

Das ISD sammelte insgesamt 883 Beiträge von 92 Konten, die AIGC (Bilder, Videos, Audio) enthielten und von rechtsextremen Akteur*innen zwischen dem 13. April 2023 und dem 18. November 2024 auf Facebook, Instagram, X (früher Twitter), TikTok und YouTube veröffentlicht wurden. Die Beiträge und Kommentare wurden manuell identifiziert und gesammelt, beginnend mit einer Seed-Liste offizieller AfD-Konten und Konten, die zu anderen rechtsextremen Online-Gemeinschaften gehören, die auf Deutsch oder speziell über Deutschland posten.

Die Stichprobe umfasste drei Arten von Inhalten:

1. Von der AfD oder AfD-Politiker*innen eingestellte Inhalte, die AIGC enthalten
2. Inhalte, die AIGC enthalten, die von der Neuen Rechten produziert wurden, Informationsmedien, die mit der Neuen Rechten verbunden sind, wie die Junge Freiheit und das Compact Magazin
3. Inhalte, die AIGC enthalten und von Online-Communities und Einzelpersonen geteilt wurden, die die AfD unterstützen

Erkennung von AI-generierten Inhalten (AIGC)

Die Erkennung von AIGC ist immer schwieriger geworden. Die Modelle und Werkzeuge, die zur Erstellung verwendet werden, werden ständig weiterentwickelt und verbessert, so dass die Inhalte selbst authentischer erscheinen. Es gibt jedoch mehrere Indikatoren, die vom ISD verwendet werden, um festzustellen, ob ein Inhalt KI-generiert wurde oder nicht. Dazu gehören:

- Räumliche und visuelle Unstimmigkeiten, einschließlich unterschiedlicher Rauschmuster in Videos und Farbunterschiede zwischen bearbeiteten und unbearbeiteten Bildabschnitten
- Zeitliche Unstimmigkeiten, z. B. Unstimmigkeiten zwischen Sprache und Mundbewegungen bei Videos
- Deformierte Hände und Gliedmaßen, einschließlich Hände und Gliedmaßen, die unnatürlich gelenkig sind oder nicht mit dem Rest des Körpers verbunden sind, insbesondere im Hintergrund eines Bildes
- Falsch geschriebene Wörter und verstümmelte Buchstaben, die nicht Teil eines echten Alphabets sind, auf Gegenständen, Postern und Wänden in Bildern und Videos
- Uneinheitliche Haartextur in Bildern und Videos
- Übermäßig glänzende, "gerenderte" Qualität, die eine unnatürliche Hauttextur erzeugt und Bilder und Videos übersättigt oder wie Gemälde erscheinen lässt

- Gesellschaftlich oder kulturell unwahrscheinliche Ereignisse und unrealistische Szenen, wie z. B. die Darstellung von Mitgliedern der Ampelkoalition ([3]) als Obdachlose in Bildern und Videos

Zusätzlich zu diesen Leitlinien nutzte das ISD TrueMedia.org[4], um einzelne Social-Media-Beiträge auf KI-generierte Deepfakes zu testen. Auf der Grundlage einer Reihe verschiedener KI-Detektoren analysiert TrueMedia Video, Bild und Audio. Die Detektoren untersuchten vier Kategorien:

1. Gesichtsmanipulation – unterscheidet Deepfakes von echten Gesichtern oder ob andere Methoden wie Gesichtsüberblendungen, Vertauschungen oder Nachstellungen verwendet wurden
2. Generierte KI – erkennt, ob das Bild mit gängigen Tools wie Dall-E, Stable Cascade, Stable Diffusion XL, CQD Diffusion, Kadinsky, Wuerstchen, Titan, Midjourney, Adobe Firefly, Pixart, Glide, Imagen, Bing Image Creator, LCM, Hive, Deepfloyd und jedem Generative Adversarial Network (GAN) erstellt wurde
3. Visuelles Rauschen – erkennt, ob Artefakte durch Manipulation oder Erzeugung in einem Bild vorhanden sind, einschließlich Pixel- und Farbabweichungen
4. Audio – erkennt, ob es Spuren gibt, dass Audio manipuliert oder geklont wurde.

In die Stichprobe wurden Beiträge und Kommentare aufgenommen, die eines oder mehrere der in der obigen Liste genannten Merkmale aufwiesen und/oder einen TrueMedia-Score ergaben, der auf "erhebliche Hinweise auf Manipulation" schließen ließ.

Das ISD stellte fest, dass die Schwierigkeit der Erkennung zwischen April 2023 – dem Datum des ersten Beitrags mit AIGC in der Stichprobe – und dem Zeitpunkt der Erstellung dieses Berichts (Februar 2025) zunahm, was auf erhebliche technische Verbesserungen der Qualität der Inhalte zurückzuführen ist. Zuweilen fügten die Nutzenden ästhetische Filter oder Text zu den Inhalten hinzu oder präsentierten sie als Kunstwerke. AIGC ist schwieriger zu erkennen, wenn Text in großer Schrift über die Komposition eines Videos oder Bildes gelegt wird und diese verdeckt. All diese Faktoren haben die Erkennung von AIGC weiter erschwert, so dass das Ausmaß von AIGC wahrscheinlich höher ist, als die Ergebnisse dieses Berichts zeigen.

Qualitative und quantitative Analyse

Die vom ISD ermittelten AIGC wurden qualitativ kodiert, um eine thematische Analyse zu ermöglichen und Veränderungen im Nutzerverhalten sowie die Nutzung von Funktionen der Social-Media-Plattformen zu ermitteln. Die quantitative Analyse wurde verwendet, um die potenzielle Reichweite und Verbreitung der Inhalte zu ermitteln.

[3] Ampelkoalition" ist die Bezeichnung für eine deutsche Regierungskoalition, die aus der Sozialdemokratischen Partei (SPD), der Freien Demokratischen Partei (FDP) und Bündnis 90/Die Grünen besteht, wie es bei der Erstellung dieser Studie der Fall war. Der Begriff leitet sich von den Parteifarben der Koalitionspartner ab, die mit den Farben Rot (SPD), Gelb (FDP) und Grün (Grüne) einer Ampel ähneln.

[4] TrueMedia.org hat seinen Betrieb am 14. Januar 2025 eingestellt. Alle Recherchen im Zusammenhang mit diesem Beitrag, einschließlich der Nutzung von Truemedia.org für die Analyse von KI-Inhalten, wurden vor diesem Datum durchgeführt.

Repräsentativität der Stichprobe

Mit der AfD, der Neuen Rechten, rechtsextremen Community-Gruppen und Inhaltserstellern, rechtsextremen Musikkanälen, Einzelpersonen und KI-generierten "Influencern" als Schwerpunkt der Studie wollte das ISD eine ähnliche Anzahl von Beiträgen auf jeder der fünf Plattformen erfassen. Pro Plattform wurden etwa 200 Beiträge gesammelt, mit Ausnahme von YouTube, wo das ISD nur 85 Videos mit AIGC-Beiträgen dokumentierte. Die geringere Anzahl von YouTube-Beiträgen kann durch die Tatsache erklärt werden, dass weniger AIGC-Beiträge auf der Plattform verfügbar sind, da YouTube videobasiert ist und AIGC-Videoinhalte mehr Raffinesse erfordern als bildbasierte AIGC-Inhalte. Da die ISD-Analyst*innen eine ausgewogene Stichprobe über alle Plattformen hinweg anstrebten, ist die Stichprobe möglicherweise nicht unbedingt repräsentativ für die tatsächliche Verteilung von AIGC-Inhalten über die Plattformen.

Als Ausgangspunkt für die Entdeckung der rechtsextremen AIGC auf X, Instagram und Facebook wurde eine Startliste offizieller AfD-Konten verwendet. Posts von Haupt- und lokalen AfD-Accounts sowie von Accounts einzelner AfD-Politiker*innen machten 50 Prozent (etwa 100 Posts) der Stichprobe für diese Plattformen aus. Die Stichprobe ist daher nicht repräsentativ in Bezug darauf, welche Akteur*in am aktivsten bei der Veröffentlichung von AIGC ist.

Das offizielle Konto der AfD auf Bundesebene ist seit Mai 2022 von TikTok gesperrt worden. Infolgedessen basierte die Datenerhebung auf TikTok auf Empfehlungen, die der Algorithmus über die "For You Page" aussprach. Die Analyst*innen stützten sich stärker auf Konten einzelner AfD-Politiker*innen, die in der Regel weniger AIGC enthielten als die Hauptkonten der AfD auf anderen Plattformen. Daher machten die AIGC von lokalen AfD-Konten und Konten einzelner AfD-Politiker*innen 19 Prozent aller auf TikTok gesammelten Beiträge aus.

Für YouTube basierte die Datenerhebung auf einer Seed-Liste von AfD-Konten, ähnlich wie bei Facebook, X und Instagram. Da sich AfD-Accounts jedoch im Allgemeinen eher auf bildbasierte als auf videobasierte AIGC konzentrieren, fanden wir auf YouTube deutlich weniger Inhalte von Haupt-, Lokal- und Einzelkonten von AfD-Politiker*innen als auf den anderen Plattformen (8 Prozent aller Beiträge).

Hauptakteur*innen und ihre Nutzung der generativen KI

Die Alternative für Deutschland (AfD)

Während des Beobachtungszeitraums sammelte ISD eine Stichprobe von über 350 Beiträgen, in denen die AIGC in sozialen Medienkanälen der AfD auf Facebook, Instagram, X, TikTok und YouTube erwähnt wurde.

Über alle Plattformen hinweg kamen diese Beiträge überwiegend vom AfD-Hauptpartei-Kanal und von einem Konto, das mit dem Bundestagsabgeordneten Norbert Kleinwächter verbunden ist, dessen Social-Media-Aktivitäten stark auf AIGC ausgerichtet sind. Zu den

anderen AfD-Mitgliedern, die AIGC auf ausgewählten Plattformen aktiv einsetzten, gehörten der Bundestagsabgeordnete Maximilian Krah (TikTok und YouTube) sowie der AfD-Landtagskandidat Sebastian Wippel (YouTube) und der Kommunalwahlkandidat Sven Hämisch (TikTok). Außerhalb dieser Stichprobe war der früheste AIGC, der vom Hauptkonto der AfD auf allen von der ISD erfassten Plattformen gepostet wurde, im August 2022.

Abbildung 1 zeigt die monatliche Anzahl der Posts von AfD-Accounts mit AIGC auf Facebook, Instagram, X und TikTok zwischen Mai und Oktober 2024.[5] YouTube wurde von dieser Visualisierung ausgeschlossen, da nur zwei der sieben untersuchten AIGC-Posts innerhalb des angegebenen Zeitraums veröffentlicht wurden. Im Oktober war die Aktivität auf X (26 Beiträge) und Instagram (21 Beiträge) am höchsten, während auf Facebook (4 Beiträge) und TikTok (1 Beitrag) deutlich weniger AIGC-Veröffentlichungen von AfD-Konten verzeichnet wurden. Instagram und X waren bei der monatlichen AIGC-Aktivität durchweg führend, außer im Juli und August, als Facebook den höchsten monatlichen Wert erreichte. Im September war die AIGC-Aktivität von AfD-Accounts auf allen Plattformen am geringsten, mit Ausnahme von TikTok, wo drei AIGC-Posts veröffentlicht wurden, was dem höchsten Wert im Mai entsprach. Trotz Schwankungen der Zahlen auf den einzelnen Plattformen veröffentlichten AfD-Konten zwischen Mai und Juni 2024 durchweg 43 bis 47 AIGC-Posts auf allen Plattformen.

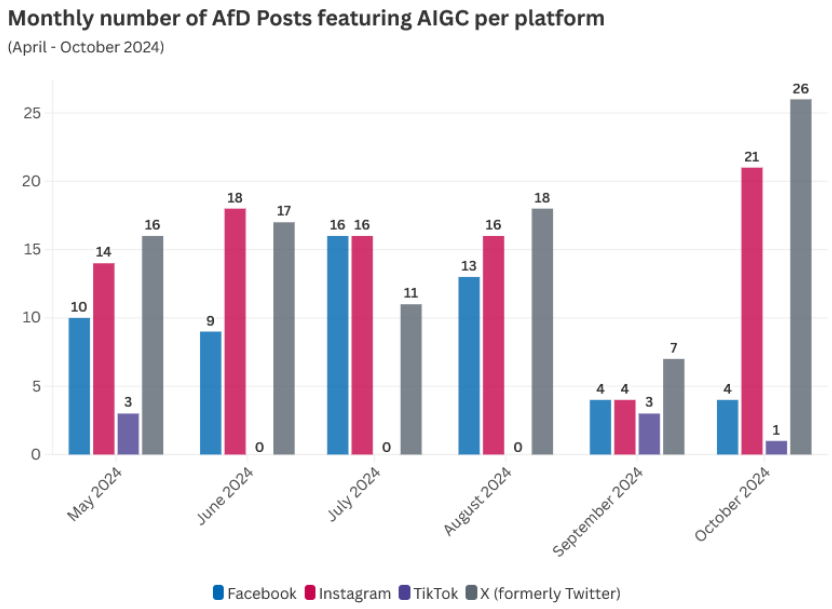


Abbildung 1: Monatliche Anzahl der AfD-Beiträge mit AIGC pro Plattform innerhalb der ISD-Stichprobe, erfasst nach Beitragsdatum.

[5] Da die in der Stichprobe erfassten Beiträge pro Akteur nicht innerhalb eines einheitlichen Datumsbereichs gesammelt wurden, wurde der Prozentsatz der Beiträge mit AIGC von der AfD ab Mai 2024 berechnet, dem frühesten Monat, in dem AIGC-Beiträge von parteieigenen AfD-Konten auf allen Plattformen einheitlich erfasst wurden.

Das ISD ermittelte den Prozentsatz der Beiträge mit AIGC, die von AfD-Konten, die von Parteien und Politiker*innen betrieben werden, auf jeder der drei untersuchten Plattformen zwischen Mai und Oktober 2024 veröffentlicht wurden. X führte die Liste mit 23 Prozent an, gefolgt von Instagram (19 Prozent) und Facebook (17 Prozent). TikTok und YouTube wurden nicht berücksichtigt, da der AfD-Hauptaccount auf TikTok seit 2022 gesperrt ist und auf den entsprechenden YouTube-Accounts der AfD nur sehr wenige Beiträge verfügbar waren.

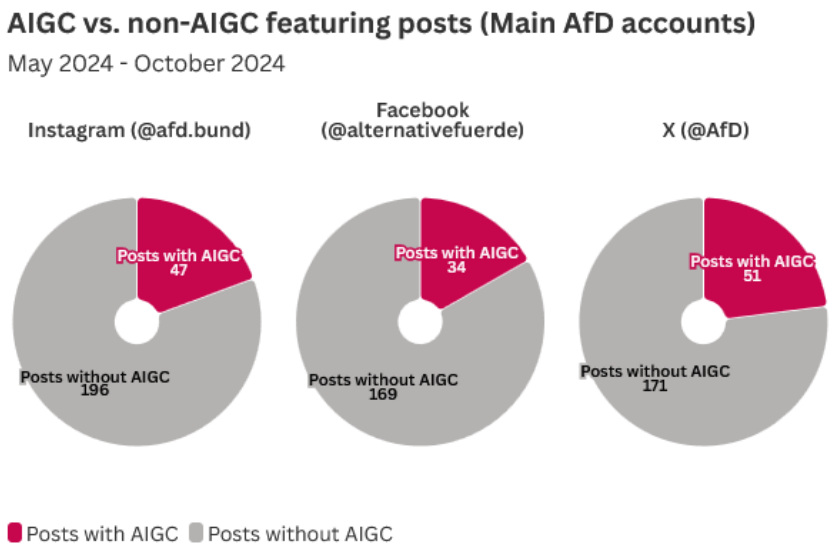


Abbildung 2: Anteil der AIGC-Beiträge an der Gesamtzahl der Beiträge auf den (von der Partei betriebenen) Hauptkonten der AfD

Die untersuchten AIGC-Beiträge der AfD verbreiteten häufig eine migrantenfeindliche Stimmung, indem sie Migrant*innen entweder als die Hauptursache für Kriminalität in Deutschland darstellten (insbesondere Messerangriffe und sexuelle Gewalt) oder sie beschuldigten, Sozialleistungen und öffentliche Dienste auszunutzen. Der Inhalt enthielt auch häufig Bilder, die zur "Remigration" aufriefen - der umfassenden Abschiebung ethnischer Minderheiten unabhängig von ihrem Einwanderungsstatus.

Das ISD fand auch regierungsfeindliche oder CDU-feindliche (Christlich-Demokratische Union) Narrative. KI wurde verwendet, um das Aussehen von Politiker*innen, ihre Kleidung und ihre Umgebung zu verändern, mit dem offensichtlichen Ziel, persönliche und/oder politische Abneigung gegen diese Gruppen zu schüren. Wie in den folgenden Abschnitten erwähnt, kamen diese Inhalte bisweilen einer gezielten Belästigung einzelner Beamt*innen gleich. Zu den weiteren Verwendungszwecken der generativen KI gehörten Beiträge, die eine klimafeindliche Stimmung fördern (indem sie Klimaaktivist*innen als Terrorist*innen darstellen), Anti-LGBTQ-Narrative+ und die Ablehnung der EU.

Obwohl Artikel 35.1.k. des Gesetzes über digitale Dienste (DSA) die Plattformen dazu auffordert, KI-generierte oder manipulierte Inhalte zu kennzeichnen, die fälschlicherweise als authentisch erscheinen könnten, wiesen über alle Plattformen hinweg weniger als 2 Prozent

der von der AfD erfassten Beiträge mit KI irgendeine Form der Kennzeichnung auf. In den Beiträgen, in denen angegeben wurde, dass sie mit Hilfe von KI erstellt wurden, nutzten die Erstellenden der Inhalte in den meisten Fällen nicht das plattformeigene System zur Kennzeichnung, sondern verwendeten stattdessen eine besonders kleine Schrift am Rand der Bilder oder Videos. Diese offensichtliche Verschleierung scheint im Widerspruch zu den öffentlichen Äußerungen von AfD-Funktionär*innen zu stehen, wie z. B. Sandro Scheer, AfD-Kreisvorsitzender von Göppingen, der sich in einem Interview mit SWR Aktuell transparent über den Einsatz von KI in der Partei äußerte.

Informationsökosystem Neue Rechte

Während der Datenerhebungsphase identifizierte das ISD Videos und Bilder mit KI-generierten Elementen, die von den deutschen Informationskanälen der Neuen Rechten (Junge Freiheit, Compact Magazin und Deutschland Kurier) veröffentlicht wurden.

- Die Junge Freiheit ist eine wöchentliche Publikation mit Inhalten für ein rechtsextremes Publikum, die von der Bundeszentrale für politische Bildung als "Flaggschiff der Neuen Rechten" bezeichnet wurde. Das Blatt veröffentlicht regelmäßig Artikel, in denen muslimische Einwander*innen als Bedrohung für die deutsche Kultur dargestellt und traditionelle Familienwerte propagiert werden. Nach der Öffnung der deutschen Grenzen für Geflüchtete im September 2015 durch die damalige Bundeskanzlerin Angela Merkel stieg die Zahl der Leser*innen deutlich an. In letzter Zeit hat sich das Blatt auf Straftaten konzentriert, die von ausländischen Bürger*innen und Islamist*innen in Deutschland begangen wurden, um eine populistische Agenda zu verfolgen.

- Das Compact Magazin wurde vom Bundesamt für Verfassungsschutz zu einer rechtsextremen Organisation erklärt. Im Juli 2024 wurde es vom deutschen Bundesministerium des Innern vorübergehend verboten, weil es zum Hass gegen Juden, Muslime und Migranten aufstachelt und die rechtsstaatliche Demokratie in Deutschland untergräbt. Das Verbot wurde später von einem Gericht wieder aufgehoben. Jürgen Elsässer, der Chefredakteur des Compact Magazins, hat über viele Jahre hinweg offen rechtsextreme politische Organisationen wie die Patriotischen Europäer gegen die Islamisierung des Abendlandes (PEGIDA), die Identitäre Bewegung und Ein Prozent unterstützt. Die Zeitschrift verbreitet rechtsextreme Narrative, antisemitische Verschwörungsnarrative und islamfeindliche Botschaften.

- Der Deutschland Kurier ist eine rechtsextreme, AfD-nahe Zeitung, die im Jahr 2017 gegründet wurde. Das Blatt bezeichnet sich selbst als "Boulevardblatt von rechts" und verfolgt eine populistische Agenda durch Kritik an der aktuellen Einwanderungspolitik, sensationslüsterne Berichterstattung über von Einwandernden begangene Straftaten, Panikmache über die aktuelle wirtschaftliche Lage in Deutschland und die Verbreitung kremlfreundlicher Positionen zum Russland-Ukraine-Krieg. Sie ruft ihre Leser*innen dazu auf, die AfD bei den anstehenden Wahlen zu unterstützen und zu wählen, und kritisiert etablierte politische Parteien wie die Sozialdemokratische Partei (SPD), Bündnis 90/Die Grünen, die Freie Demokratische Partei (FDP) und die Christlich Demokratische Union (CDU) in diffamierender Weise.

Für Medien und Informationskanäle bietet generative KI die Möglichkeit, auf einfache Weise visuelle Inhalte zu erstellen. Für Medien, die rechtsextreme Inhalte verbreiten, wie z. B. die Junge Freiheit, das Compact Magazin und der Deutschland Kurier, bietet generative KI zusätzlich die Möglichkeit, Bildmaterial über reale Ereignisse zu erstellen, das den Effekt hat, Sicherheitsbedrohungen zu fabrizieren oder zu übertreiben, um Angst und Spaltung zu schüren.

Die ISD-Analyst*innen fanden Beispiele dafür, dass die Junge Freiheit KI zur Erstellung von Videosequenzen und Bildern in Kurzvideos verwendet, die auf YouTube, Instagram und TikTok veröffentlicht wurden. Diese Sequenzen und Bilder stellen Einwander*innen als Kriminelle und als Bedrohung für Deutschland dar; andere zeigen große Menschenmengen, die gegen Einwanderung protestieren und die AfD unterstützen. Das Compact Magazin hat auch KI-generierte Titelbilder für seine YouTube-Videos verwendet, in denen farbige Menschen als Bedrohung für die Sicherheit und den Zusammenhalt dargestellt werden. Das Magazin behauptet außerdem, eine "Influencerin" namens Larissa Wagner eingestellt zu haben, bei der es sich um eine KI-generierte Figur handelt. In mehreren Online-Posts behauptet das Blatt, sie mache derzeit ein Praktikum beim Compact-Magazin und verwende KI-generierte Videosequenzen von ihr, in denen sie Interviews gebe und führe (siehe auch "KI-"Influencerinnen"). Der Deutschland Kurier hat auf seinen Social-Media-Kanälen häufig KI-generierte Bilder gepostet, um seine Berichterstattung über Migration und angebliche Straftaten von Nicht-Deutschen zu illustrieren.

Rechtsextreme Communities und Content Creator*innen

Das ISD identifizierte mehrere rechtsextreme Facebook-Seiten, die sowohl AIGC- als auch Nicht-AIGC-Inhalte teilen. Es gab mehrere Seiten, die sich anscheinend der Verbreitung rechtsextremer AIGC widmen. Zwei Seiten verbreiten Inhalte in einem Stil, der an deutsche Propaganda aus den 1920er und 1930er Jahren erinnert. Nach Angaben des Deutschlandfunks gehört eine von ihnen zum rechtsextremen Milieu und wird als einer der am häufigsten geposteten Content Creator*innen bezeichnet. ISD fand des Weiteren eine TikTok-Seite, die mit AIGC produzierte Videos imaginäre Zukunftsszenarien des Lebens in Deutschland mit der AfD an der Macht zeigen. Die Videos sind positiv und stellen die Partei als Retterin für alle gesellschaftlichen und wirtschaftlichen Herausforderungen, denen Deutschland gegenübersteht, dar.

Musik

Auf TikTok kursieren mehrere KI-generierte rechtsextreme Lieder, die Teil der AfD-Fankultur auf der Plattform sind. Rechtsextreme Konten verwenden diese Lieder als Hintergrundmusik für ihre Videos. Wenn User beim Ansehen eines Videos auf diesen Soundtrack klicken, werden sie zu weiteren rechtsextremen Inhalten weitergeleitet. Die Lieder sind oft eingängig, schaffen ein Gefühl von Identität und Zugehörigkeit und tragen letztlich dazu bei, neue Zielgruppen für rechtsextreme Inhalte zu finden und bestehende zu verstärken.

Die ISD-Analyst*innen fanden einen YouTube-Kanal mit fast 9.000 Follower*innen. Dort werden immer wieder hochglänzende, vollständig KI-generierte rechtsextreme Musikvideos veröffentlicht, die Szenen zeigen, in denen blonde, blauäugige Deutsche von Migrant*innen bedroht werden. Zu den beunruhigenden Videos gehört eines, in dem behauptet wird, dass

viele muslimische Einwander*innen, die nach Deutschland kommen, dies tun, um zu morden, und dass sie "dem Ruf des Islam folgen". Das Video zeigt KI-generierte Bilder von Kindern, die den Eindruck erwecken, sie seien Muslime und trainierten, um Messerstiche zu begehen, was impliziert, dass das Töten "in ihrer Natur" liegt (Abbildung 3). Diese fabrizierten Bilder verstärken schädliche Stereotypen, anstatt reale Personen oder Situationen abzubilden.

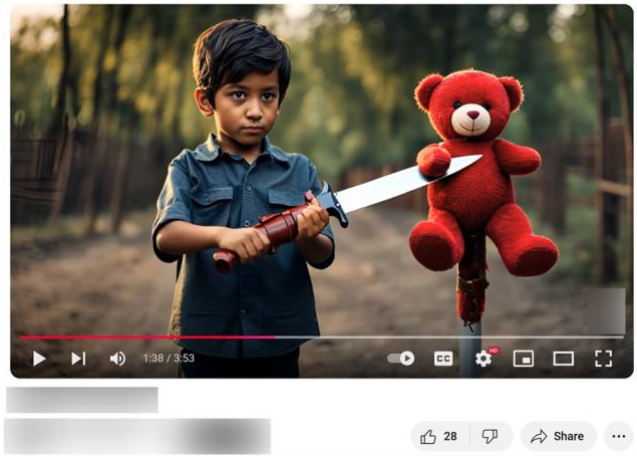


Abbildung 3: KI-generiertes Video, in dem behauptet wird, dass viele muslimische Einwanderer nach Deutschland kommen, um Morde zu begehen

Der Text des KI-generierten Musikvideos in Abbildung 3 lautet wie folgt:

"Eins, zwei, Messer, komm rüber. Drei, vier, er steht vor der Tür Deutschlands. Fünf, sechs, koste das Blut in der Nacht. Sieben, acht, gute Nacht, Deutschland ist aufgewacht."

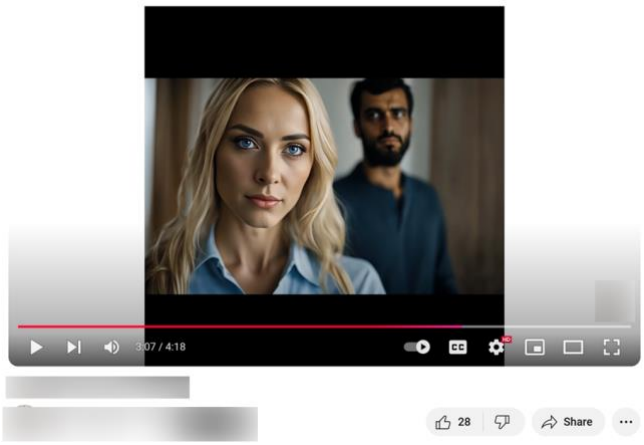


Abbildung 4: KI-generiertes Video, das Frauen davor warnt, Beziehungen mit muslimischen Männern einzugehen.

Der Text eines anderen rechtsextremen Musikvideos (Abbildung 4) warnt deutsche Frauen davor, Beziehungen mit muslimischen Männern einzugehen und greift dabei auf seit langem bestehende rassistische Tropen zurück, in denen Ausländer (insbesondere Muslime) als von Natur aus gefährlich für Frauen dargestellt werden:

"Verliebt in einen Talahon[6] , niemals, das bringt dich nur nah an den Tod. Deine Freiheit, wäg dein Leben in Gefahr, ein deutsches Mädchen gefangen im falschen Moment. Verliebt in ein Talahon, lass es sein, bevor du daran zerbrichst, er nimmt dir alles, gibt dir nichts zurück, diese Liebe ist nur Lüge, nur ein Trick. Unterdrückung einer Frau, das ist sein Glück."

Begleitet werden diese Texte von einem KI-generierten Video, das eine blonde, blauäugige deutsche Frau zeigt, die von einem muslimischen Mann emotional und körperlich missbraucht wird, und das davor warnt, dass Beziehungen zu Muslimen mit dem Verlust der Freiheit, Herzschmerz und Tod enden werden.

Diese Videos erhielten sowohl auf TikTok als auch auf YouTube geringe Engagement (in der Regel zwischen 10 und 500 Likes und Shares). Allerdings erreichten 6 der 103 TikTok-Videos im Datensatz mehr als 10k Likes und wurden mehr als 500-mal geteilt. Alle diese Videos riefen die Nutzer auf, die AfD zu unterstützen, um "Deutschland zu retten". Auf YouTube erreichten 5 von 9 Videos mehr als 1.000 Aufrufe; 2 der Videos wurden mehr als 100-mal geliked.

Einzelpersonen

Das ISD fand Beispiele für rechtsextremen AIGC, der von Einzelpersonen ohne eindeutige offizielle Zugehörigkeit zu rechtsextremen Organisationen oder Kanälen gepostet wurde. Die Inhalte wurden in öffentlichen und privaten Facebook-Gruppen von Anhänger*innen der AfD und Teilen der deutschen extremen Rechten sowie auf den Facebook-Seiten von einwanderungsfeindlichen Online-Communities geteilt.

Die beiden häufigsten Arten von AIGC unter Einzelpersonen waren Bilder, die zur "Remigration"[7] von Asylbewerber*innen und Einwander*innen aufriefen, und Bilder, die afrikanische Einwander*innen als faul und als Nutznießende der deutschen Sozialhilfe darstellten. Die Abbildungen 5 und 6 zeigen zwei Beispiele für diese Art von Inhalten, die häufig von anderen Nutzenden erneut hochgeladen werden. Während der ursprüngliche Beitrag mit dem Bild in Abbildung 5 auf Facebook 88-mal geteilt wurde, verzeichnete ein anderes Konto mit 81.000 Follower*innen, das das Bild hochgeladen hatte, 279 Shares.

[6] Talahon" ist eine abwertende Bezeichnung für junge Männer aus dem Nahen Osten, die auf TikTok und anderen Social-Media-Plattformen verwendet wird

[7] Die Zwangsdeportation von Migrantengemeinschaften mit der Absicht, eine ethnisch oder kulturell homogene Gesellschaft zu schaffen, im Wesentlichen eine gewaltfreie Form der ethnischen Säuberung

Abbildung 5: "Heimreise statt Einreise!" KI-generiertes Bild, das ursprünglich von der Zeitung Deutschland Kurier geteilt wurde, die typischerweise die AfD unterstützt. Das Bild wird häufig von Einzelpersonen auf verschiedenen Plattformen geteilt. "Heimreise statt Einreise!" ist ein Slogan, der in den 2010er Jahren häufig von der Nationaldemokratischen Partei (NPD) verwendet wurde, was wiederum die Verbindung zu den Rechtsextremen zeigt

Abbildung 6: "Keiner der Ampel-Politiker konnte sich die Haushaltslücke logisch erklären!". KI-generierte Bilder, die farbige Menschen als faul und vom deutschen Sozialstaat lebend darstellen, sind die zweite Art der am häufigsten von Einzelpersonen geteilten rechtsextremen AIGC

Das ISD stellte außerdem fest, dass einzelne Konten AIGC zusammen mit den Hashtags #vernetzungstweet, #vernetzungsschiff und #vernetzungszug auf X und Instagram verwendeten, eine besondere Taktik, die häufig mit rechtsextremen AIGC gekoppelt ist und möglicherweise von koordinierten Netzwerken inauthentischer Konten verwendet wird. Im Wesentlichen imitiert diese Taktik den Trend #FollowforFollow oder #FollowerTrain. Diese Hashtags dienen dazu, die Follower-Basis zu vergrößern, in der Regel kurz nach der Erstellung eines Kontos. Das ISD hat beobachtet, dass "Vernetzungstweets" vor allem von rechtsextremen Konten verwendet werden, die ihre Follower*innen - die oft als "Patrioten" bezeichnet werden - auffordern, den Inhalt zu mögen, zu kommentieren und erneut zu posten, um die Hashtags zum Trend zu machen und die Sichtbarkeit ihrer Botschaften zu erhöhen. Das ISD stellte fest, dass "Vernetzungstweets" vor allem von rechtsextremen Konten verwendet werden, die ihre Anhänger*innen - oft als "Patrioten" bezeichnet - auffordern, den Inhalt zu mögen, zu kommentieren und erneut zu posten, um die Hashtags zum Trend zu machen und die Sichtbarkeit ihrer Botschaften zu erhöhen

Die in den "Vernetzungstweets" verwendeten KI-Bilder, die etwa 6 Prozent der gesammelten Daten (n=200) ausmachen, reproduzieren hauptsächlich Narrative, die ein auf traditionellen Werten basierendes Gemeinschaftsgefühl und die Notwendigkeit, sich dem Kampf zur Rettung Deutschlands vor dem politischen Establishment anzuschließen, fördern. Einige dieser Beiträge erhielten bis zu 6.600 Likes, 536 Kommentare und 2.500 Shares, was ihnen Potenzial für eine weite Verbreitung gibt.

Abbildung 7: "Lasst uns einen #Vernetzungstweet starten! Wer ist dabei?": Posting mit #Vernetzungstweet und dem KI-generierten Bild eines blauen Zuges mit dem AfD-Logo und der deutschen Flagge, bearbeitet, um die Ästhetik eines Gemäldes zu erreichen.

KI-generierte "Influencerinnen"

Bei der Datenerhebung fand das ISD drei Profile von KI-generierten rechtsextremen "Influencerinnen". Darunter befinden sich zwei Konten, die vorgeben, "echte Menschen" zu sein, und ein Konto, das eindeutig eine fiktive Figur ist:

1. "Larissa Wagner", eine KI generierte "Influencerin", behauptet, eine 22-jährige Christin aus Senftenberg in Brandenburg zu sein. Auf dem Account werden Porträts einer jungen Frau geteilt, begleitet von politischen Aussagen, in denen sie die AfD unterstützt, die deutsche "Ampel"-Koalitionsregierung kritisiert und Gender Mainstreaming angreift. Larissa Wagner ist auf Instagram und X aktiv. Wie bereits erwähnt, kündigte das rechtsextreme Magazin Compact an, dass Larissa Wagner als Praktikantin eingestellt wird und eine eigene Kolumne erhält.
2. "Sophias_world" ist ein Konto mit einem KI-Avatar, der sich als junge Frau ausgibt. Sie ist auf X aktiv, wo sie ihre Unterstützung für die AfD, Rechtsextremismus und den Kreml zum Ausdruck bringt und Bilder von sich teilt, die in einem ethnonationalistischen Stil erstellt wurden.
3. "Lara, die blonde Rebellin" ist eine KI-Instagram-Persona, die auf einem fiktiven 16-jährigen Mädchen aus einem rechtsextremen Coming-of-Age-Roman basiert. Die Figur nimmt Bezug auf den Roman, auf dem sie basiert, und wirbt für ihn, teilt Bilder und Videos mit rechtsextremen Botschaften und kommentiert Ereignisse wie den Anschlag auf den Weihnachtsmarkt in Magdeburg am 20. Dezember 2024.

Diese drei KI-generierten "Influencerinnen" weisen starke visuelle und verhaltensbezogene Gemeinsamkeiten auf: Sie alle stellen junge, attraktive, starke, gesunde Frauen dar, die dem rechtsextremen Stereotyp der "idealen" deutschen Frau entsprechen. Sie verbreiten rechtsextreme Botschaften auf subtile Weise und kopieren häufig Taktiken, die von nicht KI-generierten weiblichen rechtsextremen "Influencerinnen" angewandt werden, einschließlich des Aufbaus parasozialer Beziehungen durch das Teilen persönlicher Informationen und die Schaffung eines falschen Gefühls von Intimität sowie die Thematisierung von Bedenken wie körperliche Sicherheit von Frauen bei ihrem Publikum.

Larissa Wagner verlinkt auf ihrem Profil insbesondere auf die Accounts von rechten und Neue-Rechte-Informationsanbietern wie Compact Magazin, Junge Freiheit, Heimat Kurier und Info-DIREKT. Zuletzt trat sie in Deepfake-Videos auf dem YouTube-Kanal von Compact auf und führte für das Magazin Reportagen und Interviews durch. Sophia's Welt teilt Beiträge von AfD-Politiker*innen wie der Parteivorsitzenden Alice Weidel, Maximilian Krah und Stephan Protschka sowie Reposts von rechtsextremen Community-Seiten.

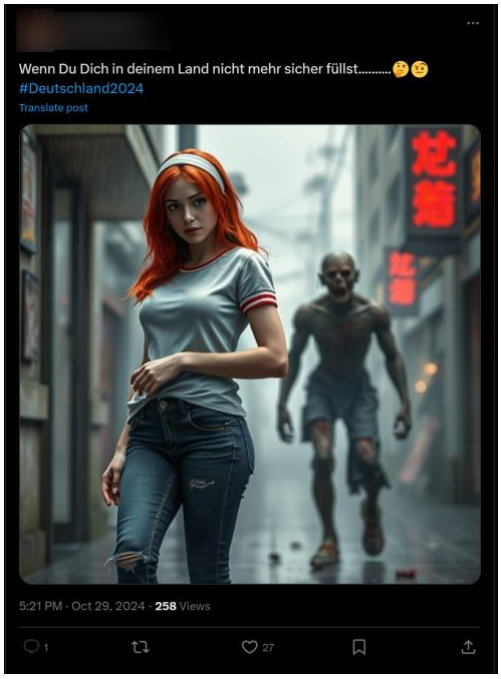


Abbildung 8: KI-generierter rechtsextreme "Influencerin" Sophia: "Wenn Du Dich in Deinem Land nicht mehr sicher fühlst... #Deutschland2024"

Im Allgemeinen erhalten die Konten sehr wenig authentisches Engagement. Roboterstimmen, Unstimmigkeiten zwischen Sprache und Mundbewegungen, glänzendes Aussehen, wenig Abwechslung bei Posen und Hintergründen sind Hinweise darauf, dass die "Influencerinnen" von KI generiert wurden. Die Beiträge erhielten selten mehr als 10 Kommentare. Im Fall von Larissa Wagner hat ihre Zusammenarbeit mit Compact im Vergleich zu den beiden anderen KI-"Influencerinnen" nicht zu einem höheren Engagement geführt. Auf allen drei Profilen kommentieren andere Nutzer*innen häufig, dass die "Influencerin" mit Hilfe von KI generiert wurde, und manche machen sich sogar über ihre Inhalte lustig. Das ISD hat jedoch auch beobachtet, dass User die Konten verteidigen und erklären, dass es keine Rolle spielt, dass die Person mit generativer KI erstellt wurde und dass "die Botschaft zählt".

Abbildung 9: KI-generierte "Influencerin" Larissa Wagner

Themen und Narrative von KI-generierten Inhalten

Das ISD kodierte die Beiträge und Kommentare qualitativ nach ihren vorherrschenden Narrativen und entdeckte dabei 14 verschiedene Narrative, die durch den Einsatz von generativer KI in dieser Stichprobe verbreitet wurden. Diese Narrative wurden danach gruppiert, ob sie bestimmte Gruppen, Parteien oder Einzelpersonen angriffen, kritisierten oder dämonisierten oder ob sie bestimmte Werte verherrlichten ("angreifende Narrative").

Art der Erzählung 1. Angreifende Narrative	Erzählung	Erläuterung
Angriffe auf Geflüchtete und Migrant*innen	RE_MIGR	Forderungen nach Massenabschiebung von Migrant*innen (euphemistisch als "Remigration" bezeichnet)
	CRIME_MIGR	Migrant*innen für Kriminalität in Deutschland verantwortlich machen, insbesondere für sexuelle Gewalt und Messerangriffe
	EXPLOIT_MIGR	Vorwürfe, dass Migrant*innen das deutsche Sozialsystem und öffentliche Einrichtungen ausnutzen
Angriffe auf andere politische Parteien und Akteur*innen	ANTI-GOV	Angriffe auf die derzeitige Regierung, die ihr Aussehen, ihre Kleidung und ihre Umgebung in übertriebener Weise verspotten
	ANTI-CDU	Angriffe auf die CDU, vor allem wegen ihrer Einwanderungspolitik während der Regierung Merkel
	ANTI-EU	Darstellung der EU als eine ruinierte Institution, die Deutschland und seiner Souveränität schadet
Angriffe auf die LGBTQ+-Gemeinschaft, Kritik an Gender Mainstreaming	ANTI_LGBTQ_GENDER	Entmenschlichung von Mitgliedern und Verbündeten der LGBTQ+-Bewegung sowie Untergrabung des Gender Mainstreaming
Angriffe auf Klimaaktivist*innen und Maßnahmen gegen den Klimawandel	ANTI_KLIMA	Entmenschlichung von Klimaaktivist*innen und Ablehnung von Maßnahmen gegen den Klimawandel

2. Verherrlichende Narrative		
Verherrlichung von Deutschland als Nation	SAVE_DE	Glorifizierung Deutschlands als idealisierte, starke Nation, die derzeit schwach ist und gerettet werden muss
Verherrlichung blonder und blauäugiger Deutscher	PHYSIQUE_DE	Die Verehrung blonder und blauäugiger Deutscher, die ein idealisiertes Bild junger, starker deutscher Männer und Frauen mit blondem Haar und blauen Augen zeichnen
Verherrlichung deutscher Traditionen	TRAD_DE	Verherrlichung traditioneller Familienwerte und der germanischen Mythologie
Verherrlichung des Kampfes für die Freiheit	FIG_FREEDOM	Narrative, die den Kampf für Freiheit und freie Meinungsäußerung in einer zensurkritischen Weise verherrlichen
3. Andere Narrative		
Unterstützung der AfD	SUP_AFD	Beiträge, in denen die AfD als Retterin Deutschlands dargestellt wird, Beiträge, in denen die AfD unterstützt wird und in denen dazu aufgerufen wird, die Partei zu wählen
Kremlfreundliche Inhalte	PRO_KREM	Werbung für die Bedeutung der deutsch-russischen Beziehungen und das Teilen kremlfreundlicher Positionen, wie die Kritik an der militärischen Unterstützung der Ukraine
Andere	OTHER	Andere Arten von Narrativen wie die wirtschaftliche Situation Deutschlands, die internationalen Beziehungen und die Proteste der Landwirte

Tabelle 1: Rechtsextreme Themen und Narrativen in von der ISD erhobenen Stichprobe von AIGC.

Abbildung 1 zeigt die Anzahl der Beiträge, die jedes in der Stichprobe gefundene Narrativ enthalten. Die fünf häufigsten Narrative waren die Verherrlichung blonder und blauäugiger Deutscher (PHYSIQUE_DE), Aufrufe zur Remigration (RE_MIGR), Verbindungen zwischen Kriminalität und Migration (CRIME_MIGR), Behauptungen, dass Migrant*innen das deutsche Sozialsystem ausnutzen (EXPLOIT_MIGR), und persönliche Angriffe auf Mitglieder der aktuellen Ampelkoalition (ANTI-GOV).

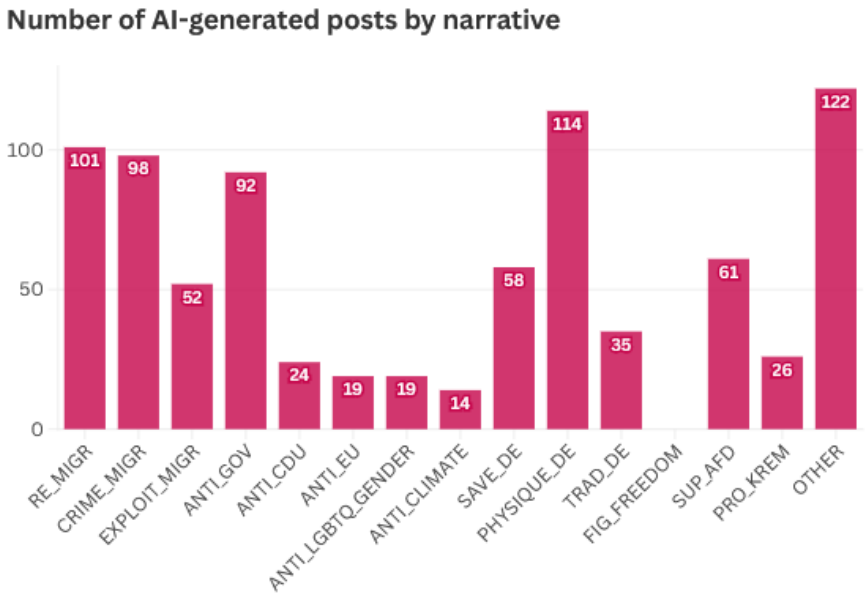


Abbildung 10: Anzahl der KI-generierten Beiträge nach Erzählung.

Eine Aufschlüsselung nach Plattformen ergab, dass die Verherrlichung von blonden und blauäugigen Deutschen (PHYSIQUE_DE) und die Unterstützung der AfD (SUP_AFD) besonders häufig auf TikTok zu finden waren. Im Gegensatz dazu waren Aufrufe zur Remigration (RE_MIGR), Beiträge, die Kriminalität und Migration in Verbindung bringen (CRIME_MIGR) und persönliche Angriffe auf Mitglieder der aktuellen Ampelkoalition (ANTI-GOV) besonders häufig auf Facebook zu finden (Abbildung 11).

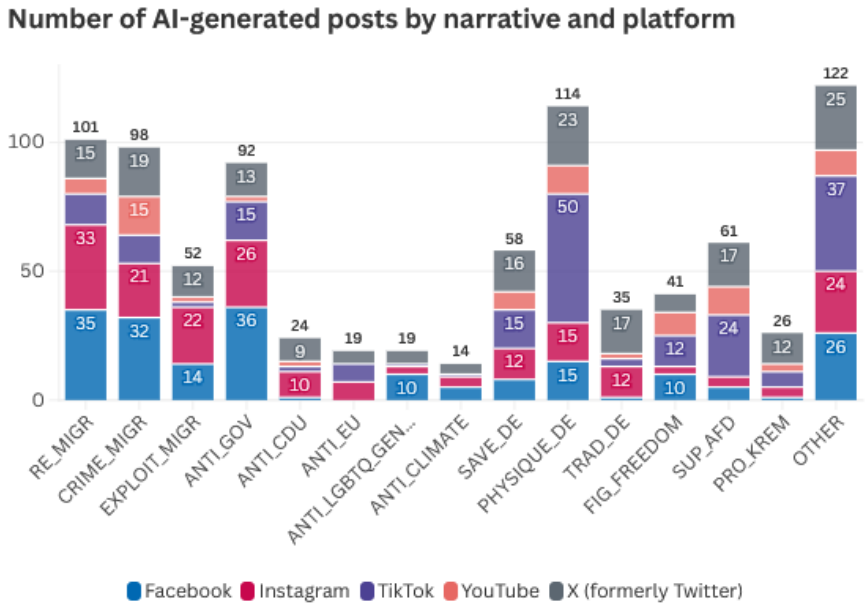


Abbildung 11: Anzahl der KI-generierten Beiträge nach Erzählung und Plattform.

Diese Divergenz zwischen den Plattformen ist angesichts ihres Formats in gewisser Weise zu erwarten: Als Kurzvideoplattform bietet TikTok rechtsextremen Akteur*innen die Möglichkeit, Narrative in einem visuellen Format zu teilen, die Unterstützung für die AfD zeigen. Auf der anderen Seite hat das ISD beobachtet, dass Facebook als bild- und textbasierte Plattform das Teilen von AIGC ermöglicht, das von Text begleitet wird und sich daher für mehr Diskussionen anbietet

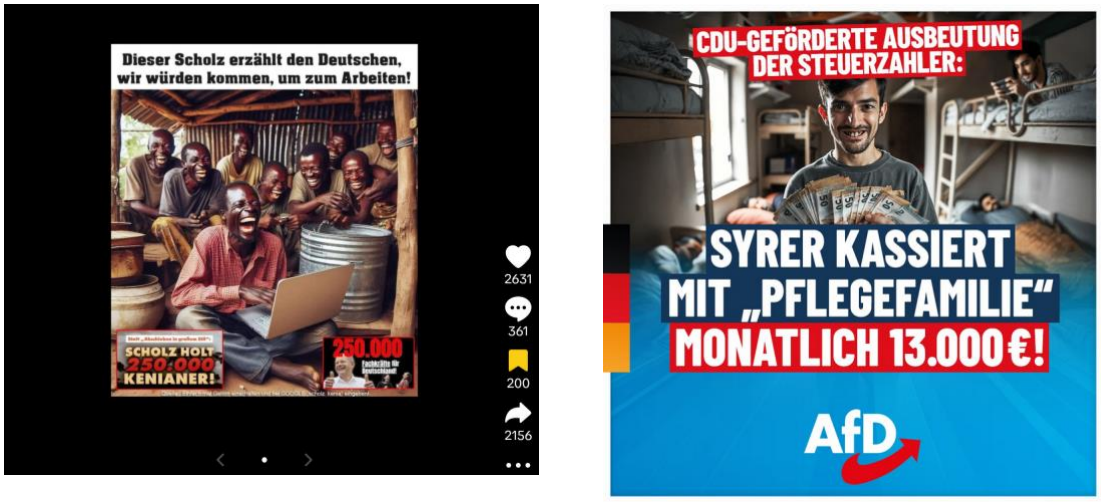
Angreifende Narrative

Das ISD stellte fest, dass rechtsextreme Online-Akteur*innen angreifende Narrative verwenden und AIGC für deren Darstellung in Bildern und kurzen Videos nutzen. Neben Original- und geteilten AfD-Posts wurden sowohl Outlets als auch einzelne Akteur*innen aus dem rechtsextremen Online-Raum beobachtet, die ebenfalls ihre eigenen AIGC mit diesen Narrativen veröffentlichten. Die Ziele der rechtsextremen "angreifenden Narrative" reichen von Geflüchteten und Migrant*innen über andere politische Parteien, Akteur*innen und Institutionen bis hin zu LGBTQ+- und Klima-Communities.

Anti-Migranten-Narrative

Geflüchtete und Migrant*innen gehörten im Vergleich zu anderen Narrativen in der ISD-Stichprobe zu den am häufigsten angesprochenen Gruppen auf allen Plattformen. Die AIGC-Inhalte der untersuchten Akteur*innen forderten häufig ihre Massenabschiebung oder "Remigration" ("RE_MIGR"; Abbildung 12). Migrant*innen wurden häufig als Täter*innen dargestellt, insbesondere bei Messerangriffen und sexueller Gewalt, neben der Forderung nach ihrer Ausweisung aus dem Land ("CRIME_MIGR"; Abbildung 13). Sie werden auch als Akteur*innen dargestellt, die das "fragile" deutsche Sozialsystem und die öffentlichen Einrichtungen ausnutzen ("EXPLOIT_MIGR"; Abbildung 14). Die rechtsextremen AIGC-Inhalte in der Stichprobe zeigten beispielsweise häufig Menschen farbiger Hautfarbe, die zufrieden lächelnd in einem Arbeitsamt sitzen und hohe Geldbeträge erhalten oder einen sorglosen Lebensstil genießen, ohne arbeiten zu müssen (siehe auch: Abbildung 15). In den Beiträgen wurde behauptet, dies habe zum wirtschaftlichen Niedergang des Landes geführt, einschließlich der Benachteiligung von Rentner*innen. Im letzteren Fall wird AIGC auch verwendet, um ältere weiße Deutsche darzustellen, die auf der Straße Pfandflaschen sammeln oder im hohen Alter im öffentlichen Dienst arbeiten.

Abbildungen 12 und 13: Abbildung 12 (links) zeigt eine farbige Person in einem Flugzeug mit der Aufschrift "Was reimt sich auf Talahon?[8] Remigration!" und fordert damit ihre Abschiebung. Abbildung 13 (rechts) ist ein Beitrag der AfD auf X, der eine ernst dreinblickende Person of Colour zeigt. Er lautet: "Berliner Polizeipräsident Slowik: Messergewalt ist jung, männlich und undeutsch!"



Abbildungen 14 und 15: In Abbildung 14 (links) ist zu lesen: "Dieser Scholz sagt den Deutschen, dass wir zur Arbeit kommen würden!" und zeigt ein KI-generiertes Bild, das auf TikTok gepostet wurde und farbige Menschen zeigt, die über einen Laptop-Bildschirm lachen.

Das KI-generierte Bild auf der rechten Seite (Abbildung 15) von einem offiziellen AfD-Konto zeigt einen lächelnden "Syrer", der einen Stapel Bargeld in der Hand hält: "CDU-geförderte Ausbeutung von Steuerzahlern: Syrer kassiert mit "Pflegefamilie" monatlich 13.000€!"

Über alle Plattformen hinweg machten diese drei migrationsfeindlichen Narrative ("RE_MIGR"; "CRIME_MIGR; "EXPLOIT_MIGR") 28 Prozent der Gesamtstichprobe aus. Bei einer getrennten Betrachtung der einzelnen Plattformen erscheinen migrationsfeindliche Narrative in 41 Prozent der Facebook-Posts, gefolgt von 38 Prozent der Instagram-Posts. Auf YouTube enthalten 27 Prozent der gesammelten Beiträge solche Narrative. X folgt mit 23

[8] Talahon" ist eine abwertende Bezeichnung für junge Männer aus dem Nahen Osten, die auf TikTok und anderen Social-Media-Plattformen verwendet wird

Prozent der Beiträge, während TikTok mit 13 Prozent der Beiträge, die sich gegen Migrant*innen richten, den geringsten Anteil aufweist.

Gegen das politische Establishment gerichtete Narrative

AIGC, die sich über deutsche Politiker*innen lustig machten, machten 15 Prozent der Gesamtstichprobe aus. Die häufigsten Ziele über alle Plattformen hinweg waren Mitglieder der Ampelkoalition, der Sozialdemokratischen Partei (SPD), der Grünen (Bündnis 90/Die Grünen) und der Freien Demokratischen Partei (FDP) (10 Prozent), gefolgt von der Christlich Demokratischen Partei (CDU) mit 3 Prozent und der Europäischen Union (EU) mit 2 Prozent. Wie bereits erwähnt, umfasst die Verhöhnung eine Veränderung oder Übertreibung der körperlichen Merkmale sowie dramatische Veränderungen der Kleidung und des Umfelds der Politiker*innen. Prominente Beispiele für Politiker*innen, die von AIGC in diesem Datensample ins Visier genommen wurden, sind Olaf Scholz und Nancy Faeser (SPD), Robert Habeck, Annalena Baerbock und Ricarda Lang (Bündnis90/Die Grünen), Christian Lindner (FDP) und Friedrich Merz (CDU).

Im Falle von Politikerinnen ist der Einsatz von KI besonders diskriminierend, da er sich stark auf die visuelle Vergrößerung von gewichts- oder altersbezogenen Merkmalen stützt (Abbildungen 17 und 18). Dies ist keineswegs ein neuer Trend, sondern deckt sich mit früheren Untersuchungen in verschiedenen Regionen, die die unverhältnismäßigen Auswirkungen von KI auf Frauen, die Einstellung der Öffentlichkeit ihnen gegenüber und die anschließenden Auswirkungen dieser Inhalte auf ihre politische Beteiligung zeigen.

Trotz ihrer ansonsten konservativen Haltung gegenüber der Zuwanderung wird die CDU von rechtsextremen Akteur*innen als die Partei dargestellt, die für einen vermeintlichen Zustrom von Migrant*innen und die "prekäre Lage" des Landes seit der Regierung Merkel verantwortlich ist.

Abbildung 16: "Heuchelei und Doppelmoral: Klima-Ampel [Koalition] fliegt mit drei Jets nach Indien!": KI-generiertes Bild, das Bundeskanzler Olaf Scholz (SPD) und Bundesaußenministerin Annalena Baerbock (Grüne) zeigt, die mit KI-Flugzeugen über das Taj Mahal fliegen

Abbildungen 17 ("Erst mal entspannen! Grüne Parteiführung tritt zurück!") & 18 ("Linksextremismus macht hässlich"): KI-generierte 'Bilder' von Ricarda Lang (Grüne) und Nancy Faeser (SPD), die sich über ihr Aussehen lustig machen

Narrative, die auf die LGBTQ+- und Klimabewegung abzielen

Zuletzt umfasste diese Stichprobe von AIGC-Inhalten auch Anti-LGBTQ+- und Anti-Klima-Stimmungen. Bei ersterem wurde beobachtet, dass AIGC genutzt wurde, um Verbündete der LGBTQ+-Bewegung zu entmenschlichen und zu verhöhnen, sowie um gegen Gender Mainstreaming zu protestieren[9] (Abbildung 20). In dieser Datenstichprobe wurde AIGC auch gegen Klima-Aktivismus eingesetzt, indem beispielsweise Aktivist*innen wie Greta Thunberg und Mitglieder der "Last Generation"-Bewegung mit Terroristen gleichgesetzt wurden (Abbildung 19) und Klimamaßnahmen der Regierung als verschwenderische Maßnahme dargestellt wurden.

[9] Hajek, K. (2020, 27. Februar). Die AfD und rechte (Anti-)Gender-Mobilisierung in Deutschland. LSE Blogs. https://blogs.lse.ac.uk/gender/2020/02/27/the-afd-and-right-wing-anti-gender-mobilisation-in-germany/.

Abbildung 19 (links), ein AfD-Posting auf X, zeigt einen bärtigen Mann mit einer Sprengstoffweste in einem Flughafen und stellt offenbar einen Klimaaktivisten der "letzten Generation" dar. Der Text lautet: "Klimaaufkleber an Flughäfen: Sicherheitsmängel sind auch eine Einladung an Islamisten"

Abbildung 20 (rechts), ein Beitrag eines Adlers mit einer deutschen Flagge, der eine Ratte mit einer LGBTQ-Flagge jagt

Verherrlichende Narrative

Die in unserer Stichprobe analysierten rechtsextremen Akteur*innen nutzten generative KI, um Narrative zu schaffen, die Deutschland als idealisierte, starke Nation darstellen, die derzeit schwach ist und gerettet werden muss. Häufig betont die AIGC die Bedeutung des "Kampfes für Freiheit und Meinungsfreiheit", während sie gleichzeitig die Demokratie in der Praxis in Deutschland und die deutschen staatlichen Institutionen angreift.

Viele der erstellten Bilder und Videos zeigen einen großen, starken und mächtigen Adler, der Deutschland repräsentiert und über die Nation wacht. In einigen der Bilder und Videos wird der Adler selbst als zu retten dargestellt (Abbildung 21).

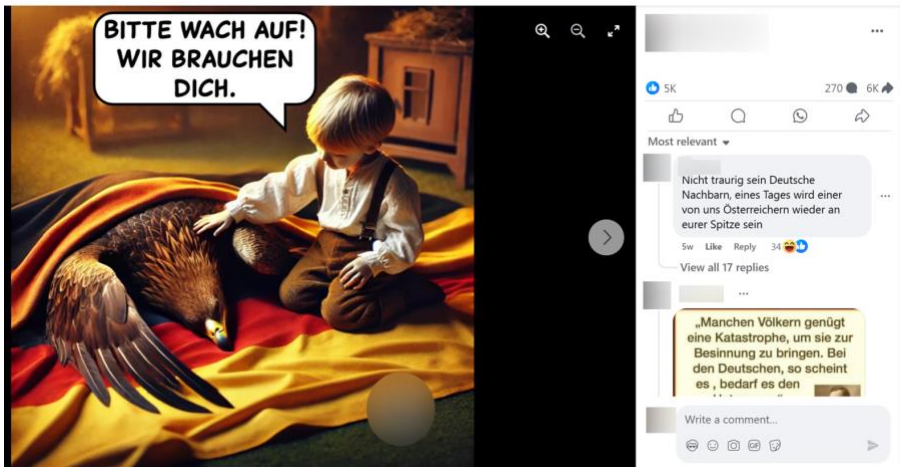


Abbildung 21: "Bitte wach auf! Wir brauchen dich.': KI-generiertes Kunstwerk, das Deutschland als einen starken und mächtigen Adler zeigt, der derzeit schwach ist und gerettet werden muss. Er wird von einem blonden Jungen in traditioneller Tracht gestreichelt, der auf einer deutschen Flagge sitzt.

Die Anwesenheit starker, blonder, blauäugiger, weißer deutscher Staatsbürger*innen ist ein weiteres gemeinsames Element rechtsextremer KI-generierter Bilder und Videos. Die Deutschen werden als körperlich starke und reine Ethnie dargestellt, die vor fremden Einflüssen gerettet werden muss - ein bekanntes rechtsextremes Narrativ. Diese Narrative zeigen Männer als hart, muskulös und kraftvoll, während Frauen "deutsche Traditionen" aufrechterhalten und sich um die Familie kümmern sollen und oft sexualisiert werden. Narrative, die Bilder von Frauen begleiten, betonen häufig die Bedrohung, die Einwanderer für diese Frauen darstellen, und implizieren, dass politische Veränderungen notwendig sind, um ihren Schutz zu gewährleisten.

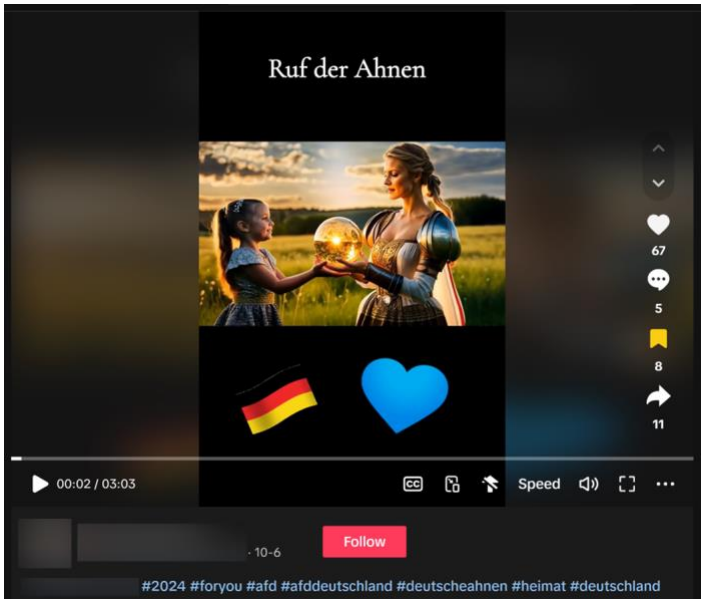


Abbildung 22: "Ruf der Ahnen": Rechtsextremes Musikvideo, das germanische Traditionen verherrlicht

Ein Konto in der Stichprobe verwendete generative KI, um Videos mit germanischer Mythologie zu erstellen, in denen die Deutschen in der Regel als ein starkes Volk dargestellt werden, das von den Germanen abstammt[10] (Abbildung 22). Die Bezugnahme auf germanische Traditionen, mächtige Führer, idealisierte Gewalt und das Überleben des Stärkeren ist eine bekannte Strategie der deutschen Rechtsextremen, um das deutsche Volk von anderen abzugrenzen und ethnische Überlegenheit zu behaupten.

Bewertung des Engagements: angreifende oder verherrlichende Narrative

Wie aus Abbildung 23 hervorgeht, erhielten angreifende Narrative auf Facebook und Instagram die meiste Zustimmung. Beiträge, die Migrant*innen als Kriminelle darstellen oder sich gegen Klimaschutzmaßnahmen aussprechen, erhielten im Durchschnitt jeweils über 5.100 Likes auf Instagram. Auch Beiträge mit Anti-CDU-Rhetorik und der Forderung nach "Remigration" erhielten im Durchschnitt 4.700 Likes. Auf allen Plattformen wurde die

[10] Die Behauptung, die Deutschen seien Nachfahren der Teutonen, wird von Rechtsextremisten instrumentalisiert, um rechtsextreme Konzepte wie Nationalismus, Autoritarismus, Ethnopluralismus und Überlegenheit gegenüber anderen ethnischen Gruppen zu illustrieren und zu bestätigen.

"Remigrations"-Rhetorik durchweg gut geteilt: Im Durchschnitt erreichten die Beiträge rund 5.200 Likes auf TikTok und 3.700 Likes auf Instagram.

Auf YouTube hatten Narrative über Migrant*innen, die das deutsche Sozialsystem ausnutzen, und Anti-CDU-Inhalte einen deutlich höheren Like-Durchschnitt als auf anderen Plattformen. Diese Spitzenwerte wurden jedoch von zwei besonders beliebten Videos von Accounts aus dem Informationsökosystem der Neuen Rechten verursacht, die die Daten erheblich verzerrten, da diese Narrative auf der Plattform normalerweise weit weniger Likes erhielten.

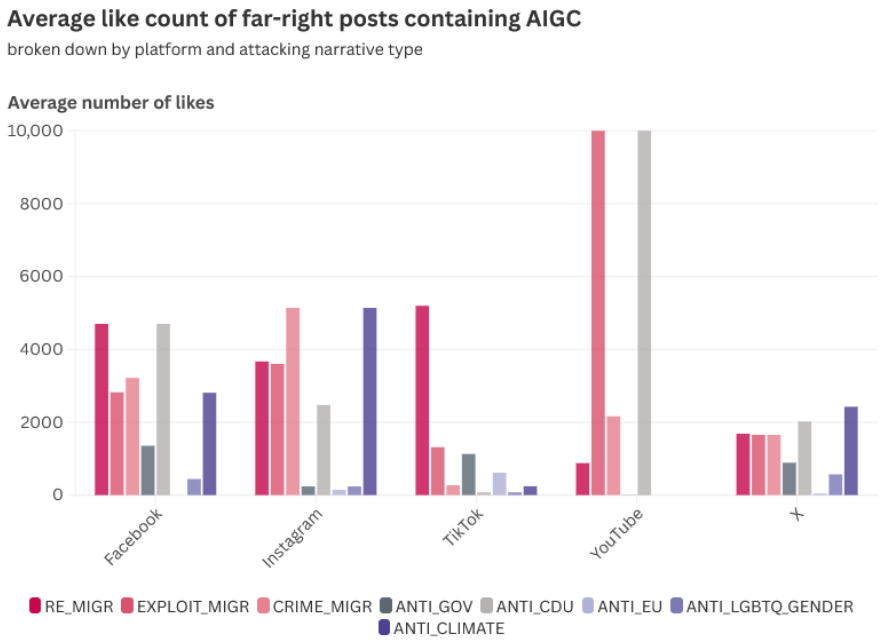


Abbildung 23: Durchschnittliche Anzahl der Likes für rechtsextreme Beiträge, die AIGC enthalten, aufgeschlüsselt nach Plattform und Art der angreifenden Erzählung.

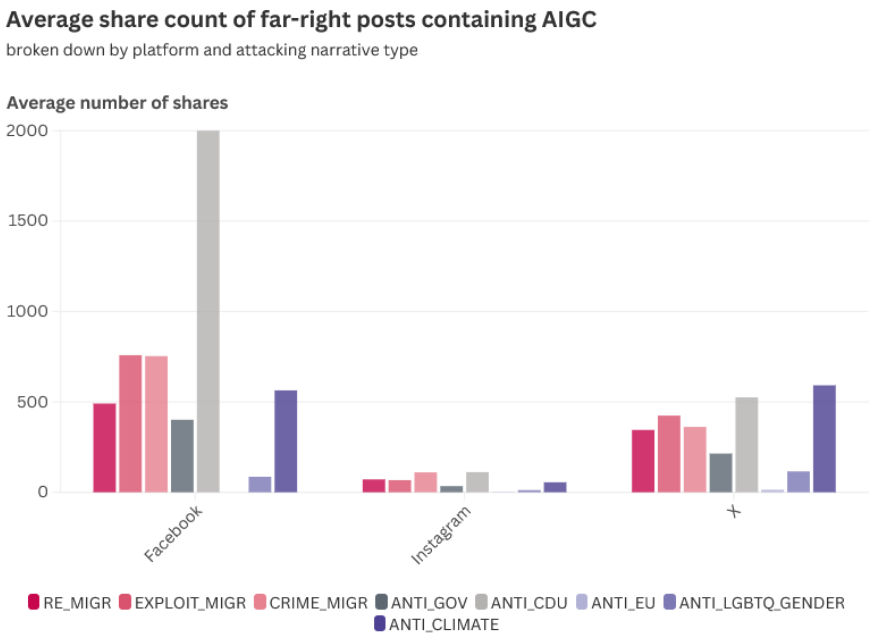


Abbildung 24: Durchschnittliche Anzahl von Anteilen rechtsextremer Beiträge, die AIGC enthalten, aufgeschlüsselt nach Plattform und Art der angreifenden Erzählung.

Abbildung 24 zeigt, dass Facebook die Plattform mit der höchsten durchschnittlichen Anzahl an geteilten Beiträgen ist. Mit Ausnahme der Beiträge, die sich gegen die EU und die LGBTQ+-Bewegung richten, wurden alle Beiträge im Durchschnitt zwischen 401 und 1.200-mal geteilt. Die Zahl der gegen die CDU gerichteten Inhalte auf Facebook sollte als Ausreißer betrachtet werden, da sie nur einen Beitrag umfasst, der vergleichsweise häufig geteilt wurde. Auf der nächstgrößeren Plattform, X, war ein ähnlicher Trend zu beobachten, allerdings waren "ANTI_EU"- und "ANTI_LGBTQ+"-Beiträge deutlich weniger beliebt als andere Kategorien.

Wie Abbildung 25 zeigt, erzielten verherrlichende Beiträge auf TikTok das meiste Engagement (in Bezug auf die Anzahl der Likes), wobei die Beiträge, die zur Rettung Deutschlands aufriefen, mit durchschnittlich rund 9.500 Likes am beliebtesten waren. Die anderen glorifizierenden Beiträge auf TikTok erreichten im Durchschnitt etwa 2.000 Likes.

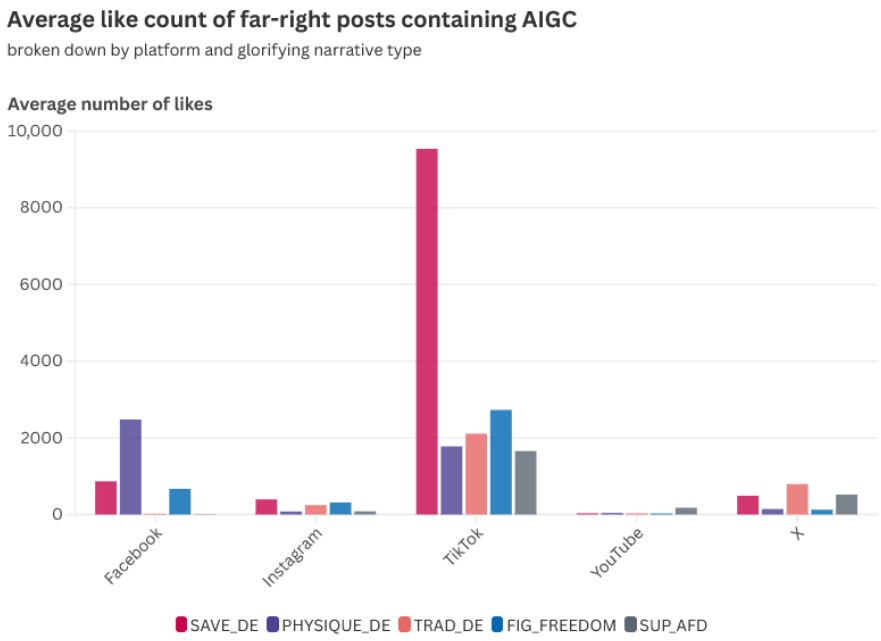


Abbildung 25: Durchschnittliche Anzahl der Likes von KI-generierten Beiträgen mit glorifizierenden Narrativen auf jeder Plattform

Das Engagement für die anderen glorifizierenden Narrative war deutlich geringer und lag in der Regel unter 1.000 Likes. Eine Ausnahme bildeten die Narrative, die blonde und blauäugige Deutsche auf Facebook glorifizierten und durchschnittlich fast 2.500 Likes erreichten.

Reaktion der Plattform und Auswirkungen auf die Politik

Schaffung und Kennzeichnung von AIGC durch den EU AI Act

Bei der Erstellung und Verbreitung von KI-generierten politischen Inhalten scheinen rechtsextreme Akteur*innen das Fehlen einer wirksamen Durchsetzung der geltenden Rechtsvorschriften durch die Plattformen auszunutzen. Dieser Abschnitt konzentriert sich auf die Politik auf EU-Ebene, wie sie für Deutschland gilt, und auf die nationale Gesetzgebung.

Als erster umfassender Rechtsrahmen für KI legt der KI-Gesetzentwurf der EU (2024) Regeln für den Umgang mit KI-bezogenen Risiken fest und bestimmt Anforderungen und Pflichten für Entwickler*innen und Anwender*innen[11] auf der Grundlage des Risikos, das mit bestimmten KI-Anwendungen verbunden ist. Ziel ist es, die "Entwicklung innovativer und verantwortungsvoller KI in der EU" zu unterstützen und die Sicherheit und die Grundrechte

[11] Im KI-Gesetz der EU wird der Begriff "Bereitsteller" wie folgt definiert: "Ein 'Bereitsteller' ist jede natürliche oder juristische Person, Behörde, Einrichtung oder sonstige Stelle, die ein KI-System unter ihrer Aufsicht nutzt, es sei denn, das KI-System wird im Rahmen einer persönlichen, nicht beruflichen Tätigkeit verwendet."

von Menschen und Unternehmen zu gewährleisten. Dies soll erreicht werden, indem sichergestellt wird, dass KI-Systeme die Grundrechte, die Sicherheit und die ethischen Grundsätze respektieren und die Risiken wirkungsvoller KI-Modelle angegangen werden. Sie verfolgt einen risikobasierten Ansatz und kategorisiert KI in vier Stufen: unannehmbares Risiko (verbotene Anwendungen wie Social Scoring); hohes Risiko (strenge Vorschriften für KI in Bereichen wie Gesundheitswesen und Strafverfolgung); begrenztes Risiko (Transparenzverpflichtungen für Chatbots und Deepfakes); und minimales Risiko (keine Beschränkungen für die meisten KI-Systeme wie Videospiele oder Spam-Filter).

Das KI-Gesetz erwähnt jedoch weder politische Parteien noch einzelne Politiker*innen als Schöpfende von KI-generierten Inhalten und auch nicht als Vertreibende. Es überträgt die Verantwortung für die Einhaltung des KI-Gesetzes den Anbieter*innen und Betreiber*innen von KI-Systemen, nicht den Nutzer*innen. KI-Systeme, die für politische Kampagnen, einschließlich der Beeinflussung von Wahlen, eingesetzt werden, werden nicht ausdrücklich als hochriskant eingestuft. Wird ein KI-System jedoch in einer Weise eingesetzt, die sich erheblich auf die Grundrechte, Wahlprozesse oder die Manipulation von Wähler*innen auswirken könnte, könnte es dennoch unter die Transparenzverpflichtungen oder eine künftige regulatorische Prüfung fallen. Die Nutzung anderer kommerziell verfügbarer KI-Anwendungen wie ChatGPT, Gemini, DallE3 oder Midjourney zur Erstellung von KI-generierten politischen Inhalten, auch durch Akteur*innen des gesamten politischen Spektrums, wird nicht ausdrücklich erfasst.

Das KI-Gesetz schreibt vor, dass außerdem Akteur*innen, die "Bild-, Audio- oder Videoinhalte erzeugen oder manipulieren, die bestehenden Personen, Gegenständen, Orten, Einrichtungen oder Ereignissen merklich ähneln und einer Person fälschlicherweise als authentisch oder wahrheitsgetreu erscheinen würden (Deep Fake), [klar und deutlich] offenlegen müssen, dass der Inhalt künstlich erzeugt oder manipuliert wurde, indem sie die KI-Ausgabe entsprechend kennzeichnen und ihren künstlichen Ursprung offenlegen". Es gibt jedoch Ausnahmen für "kreative, satirische, künstlerische, fiktionale oder analoge Werke oder Programme" sowie für den Fall, dass "eine natürliche oder juristische Person die redaktionelle Verantwortung für die Veröffentlichung des Inhalts trägt". Diese Formulierung wurde von Jurist*innen kritisiert, da sie eine subjektive Auslegung dessen zulässt, was ein Deepfake und was ein Kunstwerk ist.

Plattform-Verordnung

Rechtsrahmen wie der Digital Services Act (DSA), der AI Act und die Verordnung über terroristische Online-Inhalte (TCO) regeln in mehr oder weniger großem Umfang die Verbreitung von KI-Inhalten auf Social-Media-Plattformen. Das Gewicht der Durchsetzung und der Aufsicht über schädliche oder nicht gekennzeichnete KI-Inhalte liegt jedoch weitgehend bei den Plattformen selbst. So verpflichtet Artikel 35(1) des DSA zur Risikominderung insbesondere sehr große Online-Plattformen und Suchmaschinen (VLOPSEs) zur Durchsetzung der "Allgemeinen Geschäftsbedingungen und deren Durchsetzung" sowie zur Minderung der in Artikel 34 genannten systemischen Risiken (z. B. negative Auswirkungen auf den zivilgesellschaftlichen Diskurs und Wahlprozesse). Der DSA definiert jedoch keine Schwellenwerte, ab denen ein Risiko als "systemisch" gilt. Während die Zuständigkeiten der TCO es den zuständigen nationalen Behörden erlauben, Plattformen aufzufordern, AIGC mit

terroristischen Elementen zu entfernen, gilt dieser Anwendungsbereich nicht für andere Arten von schädlichen Inhalten, die nicht terroristischer Natur sind.

Umgekehrt stuft der DSA "schädliche Inhalte" nicht ausdrücklich als regulatorische Kategorie ein, sondern verlangt von VLOPSEs, systemische Risiken zu bewerten und zu mindern, einschließlich der Verbreitung illegaler Inhalte auf ihren Plattformen und der potenziellen Schäden durch algorithmisch verstärkte Desinformation, Hassreden oder AIGC, die die Grundrechte beeinträchtigen oder geschützte Gruppen angreifen können (Artikel 34 und 35). Konkret versteht Artikel 35 Absatz 1 Buchstabe k des DSA AIGC als *"Informationen, unabhängig davon, ob es sich um ein generiertes oder manipuliertes Bild, Audio oder Video handelt, das bestehenden Personen, Gegenständen, Orten oder anderen Einrichtungen oder Ereignissen merklich ähnelt und einer Person fälschlicherweise als authentisch oder wahrheitsgemäß erscheint*". In demselben Artikel wird darauf hingewiesen, dass eine von den Plattformen ergriffene Abhilfemaßnahme darin bestehen könnte (aber nicht muss), "sicherzustellen", dass solche Informationen *"[...] durch auffällige Markierungen erkennbar sind, wenn sie auf ihren Online-Schnittstellen präsentiert werden, und [...] eine einfach zu bedienende Funktion bereitzustellen, die es den Empfängern des Dienstes ermöglicht, solche Informationen anzuzeigen"*. Darüber hinaus sieht die DSGVO Transparenzverpflichtungen für VLOPSEs vor, die für die AIGC relevant sind, einschließlich der Artikel Artikel 14 und 17 (Transparenz bei der Moderation von Inhalten), 34 und 35 (Risikobewertungen und Maßnahmen zur Risikominderung).

Die Europäische Kommission und die nationalen Koordinatoren für digitale Dienste (DSC) üben die regulatorische Aufsicht über den DSA und die verschiedenen damit verbundenen Mechanismen wie Verhaltenskodizes und Leitlinien aus, wobei die Maßnahmen in unterschiedlichem Maße freiwillig und verbindlich sind. Für AIGC speziell ist der Verhaltenskodex für Desinformation (2022) am wichtigsten. Zum Zeitpunkt der Verfassung dieses Artikels haben alle untersuchten Plattformen, mit Ausnahme von X, den Kodex unterzeichnet.

Schließlich sind die Leitlinien für Anbieter von VLOPs und VLOSEs zur Minderung systemischer Risiken bei Wahlen (April 2024) ein wichtiges Instrument bei der Untersuchung von AIGC in sozialen Medien. Artikel 39(a) der Leitlinien fordert VLOPs und VLOSEs, "deren Dienste für die Erstellung trügerischer, voreingenommener, falscher oder irreführender generativer KI-Inhalte genutzt werden können", dazu auf, sicherzustellen, dass AIGC "unter Berücksichtigung bestehender Standards" erkannt werden kann. Es wird ausdrücklich betont, wie wichtig es ist, dies zu tun, wenn AIGC "Kandidaten, Politiker oder politische Parteien" betrifft.

Plattform Antwort an AIGC

Um die Einhaltung des DSA durch die Plattformen zu prüfen und die Reaktion der VLOPSEs auf rechtsextremes AIGC auf ihren Plattformen zu bewerten, hat das ISD 192 der auf Facebook, Instagram, YouTube und TikTok gesammelten Beiträge unter den jeweiligen Community-Richtlinien der Plattformen gemeldet. Tabelle 2 zeigt die Definition der Plattformen für verbotene AIGC:

Plattform	Definition von verbotenen KI-generierten Inhalten
Facebook	Ab Juli 2024: KI-generierte oder manipulierte Inhalte, die gegen andere Meta-Richtlinien oder Gemeinschaftsstandards verstoßen, werden verboten.
Instagram	Ab Juli 2024: KI-generierte oder manipulierte Inhalte, die gegen andere Meta-Richtlinien oder Gemeinschaftsstandards verstoßen, werden verboten.
YouTube	Technisch manipulierte oder verfälschte Inhalte, die die Nutzer*innen in die Irre führen (über dekontextualisierte Clips hinaus), z. B. um den Tod eines Regierungsvertreters zu suggerieren oder Ereignisse vorzutäuschen, bei denen ein ernsthaftes Risiko für schwere Schäden besteht. Synthetische Medien, die gegen die Community-Richtlinien von YouTube verstoßen, unabhängig davon, ob sie gekennzeichnet sind. Zum Beispiel ein synthetisch erstelltes Video, das realistische Gewalt zeigt, wenn es darauf abzielt, die Zuschauer*innen zu schockieren oder anzuwidern.
TikTok	Synthetische Medien... ... zeigen realistische Szenen, die nicht offengelegt oder gekennzeichnet sind. ... die das Bildnis (in Bild und Ton) einer realen Person enthalten, einschließlich: (1) eines Jugendlichen, (2) einer erwachsenen Privatperson und (3) einer erwachsenen Person des öffentlichen Lebens, wenn sie für politische oder kommerzielle Zwecke verwendet werden, oder wenn sie gegen eine andere Richtlinie verstoßen. ...die so bearbeitet, zusammengefügt oder kombiniert wurden (z. B. Video und Audio), dass eine Person über reale Ereignisse irregeführt werden kann. ... Verstöße gegen andere Richtlinien (Hassreden, sexuelle Ausbeutung, Belästigung,...
	Das sind die Medien: ... erheblich und täuschend verändert, manipuliert oder fabriziert wurden oder ... in irreführender Weise oder mit falschem Kontext weitergegeben werden und ... wahrscheinlich zu einer weit verbreiteten Verwirrung in öffentlichen Angelegenheiten führen, die öffentliche Sicherheit beeinträchtigen oder ernsthaften Schaden verursachen.

Tabelle 2: Definitionen der sozialen Medienplattformen für verbotene KI-generierte Inhalte

Bis zum 16. Dezember 2024 wurde keiner der gemeldeten Inhalte entfernt oder eingeschränkt, wenn er gegen die Richtlinien der Plattform verstieß (z. B. gewalttätige Inhalte), oder gekennzeichnet, wenn er irreführend sein könnte. Nur die auf TikTok gemeldeten Beiträge erhielten eine Antwort, in der es hieß, dass der Beitrag nicht gegen die Gemeinschaftsrichtlinien verstoße. Dies deckt sich mit den Ergebnissen der ISD in Bezug auf die Antworten der Plattformen bei der Meldung anderer Arten von AIGC.

Während des Forschungsprozesses stellte das ISD jedoch fest, dass 15 der von TikTok gesammelten Beiträge entfernt wurden, ebenso wie einer der von Facebook gesammelten

Beiträge. Der Grund für die Entfernung ist nicht bekannt und steht nicht unbedingt in Zusammenhang mit den KI-generierten Inhalten. Keiner der von Instagram, YouTube oder X erfassten Beiträge wurde entfernt.

Die Überprüfung des KI-Gesetzes und des DSA als Rechtsrahmen durch das ISD zeigt, dass die Plattformen die Vorschriften nicht wirksam einhalten, was böswillige Akteur*innen wie Rechtsextremist*innen ausnutzen können. Das ISD fand heraus, dass nur 3 Prozent der KI-generierten rechtsextremen Facebook-Beiträge, 2 Prozent der Instagram-Beiträge und weniger als 1 Prozent der Beiträge auf X gekennzeichnet waren. Der Prozentsatz der als KI-generiert gekennzeichneten Beiträge auf TikTok war mit 14 Prozent etwas höher, während keines der YouTube-Videos als KI-generiert gekennzeichnet war.

Die KI-Technologie wird immer fortschrittlicher; eine Studie aus dem Jahr 2024 ergab, dass die meisten Menschen nicht in der Lage sind, AIGC von echten Inhalten zu unterscheiden. Daher muss unbedingt sichergestellt werden, dass bestehende und künftige rechtliche Rahmenbedingungen die Kennzeichnung von ansonsten potenziell irreführenden AIGC durchsetzen und die Plattformen dazu veranlassen, den Kennzeichnungsprozess für die Ersteller entsprechend zu optimieren. Plattformen sollten verpflichtet werden, ihre Nutzungsbedingungen in einer Weise einzuhalten, die mit dem Schutz und der Achtung der Grundrechte, wie in der EU-Charta dargelegt, im Einklang steht, und sicherzustellen, dass die Durchsetzung wirksam gegen diskriminierende und hasserfüllte Inhalte vorgeht. Derzeit kursieren auf den Plattformen große Mengen rechtsextremer AIGC, die die Einhaltung ihrer Richtlinien zu hasserfüllten und irreführenden Inhalten in Frage stellen, insbesondere während eines Wahlkampfes. Sie werden weithin geteilt und erneut gepostet, ohne dass dies Konsequenzen für den ursprünglichen Urheber oder die Personen hat, die sie teilen und erneut hochladen.

Schlussfolgerung

Dieser Bericht hat veranschaulicht, wie generative KI für rechtsextreme Akteur*innen in Deutschland zu einem zentralen Instrument wird, um überzeugende Narrative zu schaffen, ein Gefühl der Identität und Zugehörigkeit bei ihren Anhänger*innen zu erzeugen und ihnen zu ermöglichen, neue Zielgruppen zu erreichen. Die AfD spielt eine zentrale Rolle bei der Schaffung und Verbreitung von rechtsextremen KI. Allerdings haben auch eine Reihe anderer rechtsextremer Akteur*innen – einschließlich der Neuen Rechten, Community-Seiten und rechtsextremer Inhaltsersteller wie Musikkanäle und Kunstseiten – diese Strategie übernommen.

Nach der Analyse von 883 Beiträgen mit 15 verschiedenen Arten von Narrativen hat das ISD festgestellt, dass generative KI eine Strategie ist, die von rechtsextremen Akteur*innen in Deutschland eingesetzt wird, um Narrative und Themen zu teilen und darauf aufzubauen, die in von denselben Akteur*innen erstellten Nicht-AIGC-Beiträgen zu finden sind. KI-generierte Bilder und Videos werden verwendet, um rechtsextreme Narrative ausdrucksstark zu visualisieren, die oft Angst schüren und die Nutzenden emotional ansprechen, indem sie die Bedrohung durch Migrant*innen, die Ampel-Koalition, Oppositionsparteien und LGBTQ+- und Gender-Aktivist*innen übertreiben. Zusammen mit Narrativen, die Deutschland, blonde und blauäugige Deutsche und germanische Traditionen verherrlichen, rechtsextremer Musik und

dem Aufbau von Online-Gemeinschaften hilft der Einsatz von generativer KI rechtsextremen Akteur*innen, langjährige Narrative zu fördern und Online-Gemeinschaften aufzubauen.

Generative KI bietet der AfD und anderen rechtsextremen Akteur*innen eine noch nie dagewesene Möglichkeit, ein großes Publikum mit rechtsextremen Inhalten zu erreichen, die kostengünstiger und in wesentlich kürzerer Zeit produziert werden. Die Popularität dieser Kommunikationsstrategie zeigt sich in der zunehmenden Zahl von AfD-Beiträgen, in denen AIGC verwendet wird, die entweder von der Partei selbst oder von Medienagenturen erstellt werden, die den Prozess der Inhaltserstellung weiter vereinfachen. KI hat nicht nur für die AfD, sondern auch für andere rechtsextreme Gemeinschaften und Einzelpersonen, denen eine Fülle von Programmen zur Verfügung steht, Hürden bei der Erstellung von Inhalten beseitigt. Die Tatsache, dass KI-Modelle wie DALL-E, Midjourney oder Stable Diffusion nachweislich diskriminierende Vorurteile reproduzieren und verstärken, könnte die Produktion von Inhalten für rechtsextreme Akteur*innen weiter vereinfachen.

Unsere Forschung hat ergeben, dass generative KI in bestehende Plattformtaktiken integriert wird, wie z. B. den #Vernetzungstweet auf X oder die Umgehung des TikTok-Verbots durch die AfD, indem alternative Konten verwendet werden und sich einzelne Nutzende darauf verlassen, ihre Inhalte von anderen Plattformen erneut zu teilen. In diesem Zusammenhang dient KI als leistungsstarke Ergänzung zu diesen etablierten Strategien und nicht als "Wunderwaffe" im Spielbuch der Rechtsextremen.

Die angreifenden Narrative zielen vor allem darauf ab, Migrant*innen und Farbige zu verunglimpfen, Mitglieder der Ampelregierung und der CDU aufs Korn zu nehmen und zu verhöhnen und sowohl Klima- als auch LGBTQ+-Anhänger*innen zu entmenschlichen. Das übergreifende Ziel scheint den Zielen anderer Narrative ähnlich zu sein: ein Identitätsgefühl innerhalb der rechtsextremen Gemeinschaft zu schaffen und zu fördern. Dies beruht auf der Ablehnung von Gruppen, die sie angreifen, verbunden mit der Verherrlichung "traditioneller" Werte als Antwort auf diese "Bedrohung". Generative KI ist wahrscheinlich ein äußerst wichtiges Instrument zur Erreichung dieses Ziels, da sie unbegrenzte kreative Möglichkeiten bietet, um die visuelle Wirkung solcher Narrative zu verstärken.

Eine wichtige Erkenntnis der Nutzer*innen, die sich mit allen analysierten Arten von rechtsextremen Beiträgen auseinandersetzten, ist, dass die Inhalte nicht völlig realistisch aussehen müssen, um eine überzeugende Botschaft zu vermitteln. Beiträge mit angreifenden Narrativen erhielten eine besonders hohe Anzahl von Likes, Kommentaren und Shares. In vielen Fällen lösten sie in den Kommentaren Hassreden gegen die Gruppe aus, auf die der KI-generierte Beitrag abzielte. Bei den KI-"Influencerinnen" beobachtete ISD, dass sie sehr wenig authentisches Engagement erhielten und dass sich die Nutzer*innen der sozialen Medien über sie lustig machten, weil sie von der KI generiert wurden. Allerdings gab es auch einige Nutzer*innen in den Kommentarbereichen, die erklärten, dass es für sie keine Rolle spiele, ob etwas KI-generiert sei, sondern dass im Gegenteil "die Botschaft zählt".

Dieser Bericht zeigt, dass generative KI bereits eingesetzt wird, um bestehende Polarisierungen zu verschärfen, Situationen und Umstände falsch darzustellen und Anhänger für die AfD und die rechtsextreme Szene zu gewinnen.

Rechtsextreme Akteur*innen profitieren davon, dass die Plattformen bestehende Vorschriften wie das EU-KI-Gesetz und den DSA nicht einhalten. Im Allgemeinen entfernen Social-Media-Plattformen keine KI-generierten rechtsextremen Beiträge, selbst wenn sie sich gemäß ihren eigenen Richtlinien als schädlich oder irreführend erwiesen haben. Im Superwahljahr 2024 erreichte die Debatte über den Einsatz von generativer KI und ihre Auswirkungen auf Wahlen und die Integrität demokratischer Prozesse eine neue Stufe. Generative KI wurde von Akteur*innen zur Beeinflussung von Wahlen eingesetzt, was darauf schließen lässt, dass auch der deutsche Wahlkampf und die kommende Bundestagswahl (Februar 2025) vor einer solchen Entwicklung nicht gefeit.

Empfehlungen

1. **Für die Industrie:**

- **Gewährleistung einer konsequenten und transparenten Durchsetzung der Maßnahmen gegen AIGC, Hassreden und Desinformation bei Wahlen**
    - Social-Media-Plattformen sollten ihre bereits bestehenden Richtlinien zur Inhaltsmoderation in Bezug auf AIGC, Hassrede und Desinformation bei Wahlen systematisch und vorhersehbar durchsetzen und sich dabei an den DSA-Anforderungen und den Leitlinien der Europäischen Kommission für VLOPSEs zur Minderung systemischer Risiken für Wahlprozesse ("Leitlinien der Europäischen Kommission zu Wahlrisiken") orientieren.
    - Gemäß den Leitlinien der Europäischen Kommission zu Wahlrisiken ist eine klare Kennzeichnung von AIGC unerlässlich, um die Nutzer*innen zu informieren und die Informationsintegrität zu wahren, insbesondere bei Wahlen (Absatz 39).
    - In Anlehnung an die Leitlinien der Europäischen Kommission zu Wahlrisiken sollten die Plattformen spezielle Wahlteams einrichten, die über Fachwissen in den Bereichen Inhaltsmoderation, Faktenüberprüfung, Cybersicherheit und Desinformation verfügen und lokale und sprachliche Fachkenntnisse gewährleisten.
    - Die Plattformen müssen die nationalen und EU-Gesetze in Bezug auf Informationen und Wahlintegrität einhalten, einschließlich der Fristen für das Verschweigen von Wahlen, der Transparenz politischer Werbung und der Beschränkungen für die automatische Verstärkung von Inhalten, die den öffentlichen Diskurs verzerren könnten.

- **Verstärkte Koordinierung mit Regulierungsbehörden, Hochschulen und der Zivilgesellschaft**
    - Die Plattformen sollten aktiv mit den Regulierungsbehörden der EU und Deutschlands, wie der Bundesnetzagentur, der Wissenschaft und der Zivilgesellschaft zusammenarbeiten, um politische Anpassungen und die Durchsetzung der Vorschriften, insbesondere in Bezug auf KI-generierte Desinformation bei Wahlen und Hassreden, zu unterstützen.
    - Sie sollten vor, während und nach den Wahlen Kommunikationskanäle einrichten, um aufkommende Bedrohungen zu erkennen und Reaktionen in Echtzeit zu ermöglichen, und die Koordination mit Wahlbeobachtern,

Faktenprüfern und Forschern ausbauen.

- **Durchführung und Weitergabe von internen Risikobewertungen und Maßnahmen zur Risikominderung**
    - Gemäß dem DSA müssen VLOPs und benannte Dienste systemische Risiken bewerten, die sich negativ auf Wahlprozesse und den zivilen Diskurs auswirken könnten. Die Plattformen sollten den Umfang dieser Bewertungen erweitern und sicherstellen, dass sie neu auftretende Risiken, einschließlich der Rolle von AIGC bei der Verstärkung von Wahldesinformation, gründlich analysieren. Die Ergebnisse der Maßnahmen zur Risikominderung sollten den Regulierungsbehörden der EU und Deutschlands, unabhängigen Prüfern und, wo möglich, Forschern mitgeteilt werden, um die Transparenz durch maschinenlesbare Daten zu gewährleisten.

2. **Für Regierungen und Aufsichtsbehörden:**

- **Starke Kommunikationskanäle zwischen Regulierungsbehörden und Forschenden einrichten**
    - Die Regulierungsbehörden, einschließlich der Bundesnetzagentur, sollten mit Forschenden zusammenarbeiten, um Wahlrisiken und die Wirksamkeit von Maßnahmen außerhalb von Wahlperioden zu bewerten.
    - Ein ständiger Dialog mit der Wissenschaft und der Zivilgesellschaft kann die Risikoanalyse, den Austausch von Methoden und Strategien zur Risikominderung verbessern. Plattformen und Regulierungsbehörden sollten Initiativen wie das European Network on Elections (ECNE) und die Europäische Beobachtungsstelle für digitale Medien (EDMO) unterstützen, um die Zusammenarbeit bei der Überprüfung von Fakten und der Bekämpfung von Desinformation bei Wahlen zu fördern.

ISD | Powering solutions to extremism, hate and disinformation

Amman | Berlin | London | Paris | Washington DC

www.isdglobal.org