Nr. 7

Täuschend schlecht

Generative KI in rechtsalternativen Netzwerken

Künstliche Intelligenz der generativen Art wird immer mehr zur Verbreitung politischer Inhalte eingesetzt – auch in rechtsalternativen Kontexten. Dank frei zugänglicher Bildgeneratoren schwemmen Motive mit extremistischen Inhalten den digitalen Bildermarkt. Dieser Eindruck kann zumindest entstehen, wenn Berichten über eine KI-Revolution im politischen Raum geglaubt werden darf. Unsere Studie auf Basis von Telegram-Daten zeigt ein differenzierteres Bild. Nicht einmal jedes 20. Bild basiert auf generativer KI. Die Nutzung ist divers. Allein eine Täuschungsabsicht zu unterstellen, greift hierbei zu kurz.

Unsere Empfehlungen in aller Kürze:

- Eine klar sichtbare Kennzeichnung von KI-generierten und -gestützten Bildern, so wie sie im AI Act gefordert ist, hilft Nutzenden einzuordnen, ob Bildmaterial von KI-Bildgeneratoren erstellt wurde.
- KI-generierte Bilder können »systemische Risiken« darstellen, die eine absehbare nachteilige Auswirkung auf die gesellschaftliche Debatte und Wahlprozesse haben.
- Die EU-Kommission sollte überprüfen, inwieweit Telegram nach dem Digital Services Act als sehr große Online-Plattform eingestuft werden kann. So hätte Telegram eine klare Verantwortung dafür, dieses Risiko plattformintern zu minimieren.

Datenbasis

- **Grundlage der Analyse waren 78.973 Bilder aus den Daten des hausinternen Langzeit-Monitorings rechtsalternativer Telegram-Kanäle sowie von relevanten regionalen Accounts in Brandenburg, Sachsen und Thüringen während der Hochphase des ostdeutschen Landtagswahlkampfes von 15. Juli bis 31. August 2024.**
- **Hieraus wurden nach algorithmischer Vorselektion und drei manuellen Kodierungsphasen 158 KI-generierte oder -gestützte Bilder identifiziert, bei welchen mit hoher Wahrscheinlichkeit KI-Bildgeneratoren Teil des Produktionsprozesses waren.**
- **Anhand verschiedener Kodierungskategorien zu Themen, visueller Ansprache, Missbrauchspotenzial und regionaler Herkunft wurden verschiedene Hypothesen zur Art der Gefährdung durch KI-generierte Bilder auf den Prüfstand gestellt.**

Im Vorfeld des sogenannten Superwahljahres 2024 wurde viel spekuliert, ob die mit der Veröffentlichung von ChatGPT rasant an Bedeutung gewinnende generative Künstliche Intelligenz (KI) nun eine neue Qualität der Anwendbarkeit erreicht habe. Damit einher geht die Befürchtung, dass das damit verbundene Missbrauchspotential entscheidenden Einfluss auf Wahlen und demokratische Prozesse ausüben könne.

Wie verbreitet sind KI-generierte Bilder in der Kommunikation rechtsalternativer Akteur*innen? Wie und in welchen Kontexten werden sie angewandt? Die Studie untersucht ein repräsentatives Datensample von 78.973 geteilten Bildern rechtsalternativer Akteure und von relevanten regionalen Accounts in Brandenburg, Sachsen und Thüringen während der Hochphase des Landtagswahlkampfes (15. Juli bis 31. August 2024). Dabei wurden 158 KI-generierte Bilder identifiziert, was weniger als 5 Prozent des Materials entspricht.

Zwischen Apokalypse und Normalisierung

Das rasante Wachstum von KI, einhergehend mit sinkenden Produktionskosten und steigender Qualität, erschwert Nutzer*innen, zwischen realen Fotografien und KI-generierten Grafiken zu unterscheiden. Diese Situation, in der für Nutzer*innen nicht mehr klar ist, welche Inhalte realen Ursprungs sind, eröffnet böswilligen Akteur*innen die theoretische Möglichkeit, eine Menge Schaden anzurichten – zum Beispiel durch eine falsche, aber plausible Darstellung von Personen und Szenarien, die der Desinformation dient.

Die Debatte um KI ist aber nicht nur durch die Unterschätzung von Risiken geprägt, sondern auch durch deren Überschätzung. Apokalyptische Szenarien und mediale Präsenz fördern die Verfügbarkeitsheuristik, bei der Menschen die Häufigkeit von KI-generierten Inhalten überschätzen. Dies kann die Glaubwürdigkeit von verlässlichen Informationen schwächen und demokratische Prozesse gefährden, hilft im Zweifelsfall aber vor allem denjenigen, welche glaubwürdige Informationen in Zweifel ziehen wollen: das Phänomen der Dividende des Lügners (»the liar's dividend«). Akteure können Unsicherheit nutzen, um sowohl Fälschungen glaubhaft als echt darzustellen, als auch echte Inhalte als gefälscht zu deklarieren.

Was generative KI so besonders macht, ist die Möglichkeit, ein komplett neues, synthetisches Bild zu erschaffen – basierend auf Milliarden von online zugänglichen Bildern. Der Hype um KI-Bildgeneratoren in den letzten Jahren führte vom rudimentären Craiyon zu DALL-E bis direkt zu ChatGPT. Deutliche Abstufungen bei den Bildgeneratoren sind bei Einschränkungen und Sicherheitsvorkehrungen sichtbar. Dem Gefahrenpotenzial bewusst, filtern die meisten aktuell gängigen, frei zugänglichen Bildgeneratoren explizite und irreführende Inhalte; sie können aber umgangen werden. Und Sicherheitsmaßnahmen sind nicht immer transparent oder engmaschig gefasst. So umgehen rechtsalternative Akteur*innen KI-Beschränkungen durch kreative Prompts oder unzensierte Modelle.

Gefahr oder Hype?

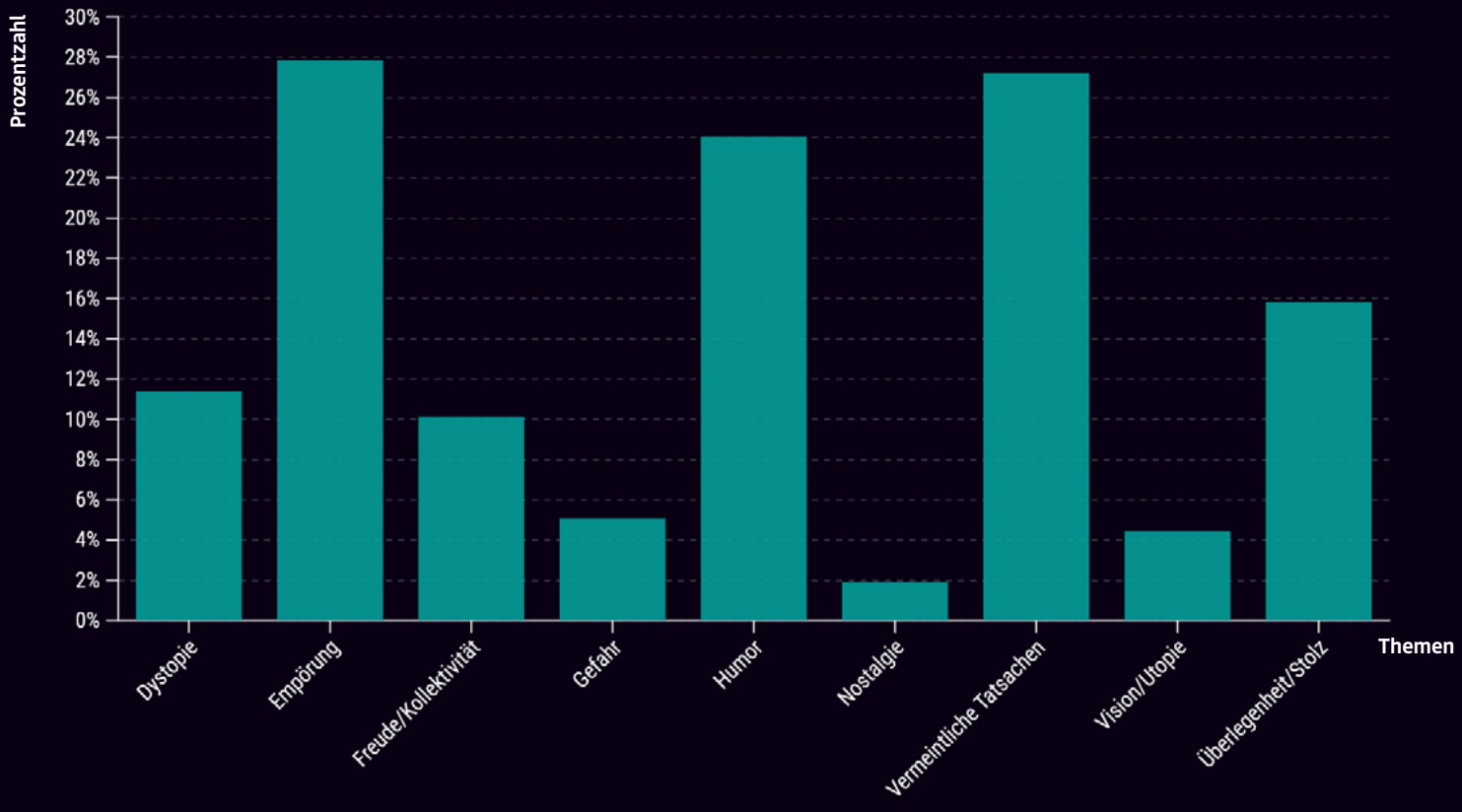
Über ein Drittel der KI-generierten und -gestützten Bilder (35,4 Prozent) im Datensatz behandeln politische Themen, das Thema Migration wurde oft mit Kriminalität verknüpft (13,2 Prozent). Hier wurde hauptsächlich mit dem diskriminierenden Bild der suggerierten Invasion, der Debatte um den Begriff »Remigration« und homogenen Menschenmassen gearbeitet. Mehr als jedes zehnte Bild enthält ein Nationalsymbol oder zeichnet eine Überhöhung der eigenen Nation – teils als weiße »Gegenkultur«. Naturmotive wiederum werden im absoluten Vergleich bei jedem fünften Bild genutzt, über die Hälfte im generischen Kontext.

Die häufigste Form der visuellen Ansprache sind „vermeintliche Tatsachen" und „Empörung", oft kombiniert mit Humor, besonders bei politischen Themen, Migration und Kriminalität. Nationalismus und Natur werden eher mit positiven Emotionen wie Freude und Nostalgie verknüpft, um zu mobilisieren oder Veranstaltungen zu bewerben. Ähnlich häufig im Datensatz zu finden sind Darstellungen von bekannten bzw. ihnen merklich ähnelnden Personen

(17,7 Prozent) und fotorealistischen Situationen (17,2 Prozent), die Elemente von Gewalt, Bedrohung oder Gefahr umfassen. Während nur ein kleiner Teil der KI-generierten Inhalte fotorealistische Darstellungen mit realen Personen kombiniert (1,9 Prozent), erfüllt fast ein Drittel der untersuchten Bilder mindestens eines dieser beiden Kriterien. Viele der fotorealistischen Bilder sind auch generischer Natur.

Trotz unseres Fokus auf die Wahlkampfhochphase in Thüringen, Sachsen und Brandenburg, stellen Bilder mit US-Bezügen einen auffällig großen Anteil (14,6 Prozent) dar. Es ist sowohl denkbar, dass deutsche Akteure diese Themen aufgreifen und bebildern als auch, dass internationales Material übernommen wird. Zudem zeigen die Ergebnisse, dass sich regelrechte „KI-Influencer" etablieren, die sich auf den Umgang mit generativer Bilder-KI spezialisiert haben und so versuchen, eine mögliche Vorreiterrolle im rechtsextremen Telegram-Kosmos einzunehmen. Außerdem nutzen rechtsalternative Akteure KI zur (Re)produktion von Memes und bekannten Meme- Charakteren.

»Ein strategischer Einsatz von generativer KI ist bisher nur in Ansätzen zu erkennen. Dass sich dies ändern wird, dafür spricht der US-Wahlkampf. Dass hiervon nicht demokratische Akteure profitieren, ist abzusehen.«



Prävalenz von KI-generierten/-unterstützten Bildern mit Bezug zu visuellen Ansprachen

Es handelt sich hier um die kondensierte Version des Themenschwerpunkts von *Machine Against the Rage*, Nr. 7 (Herbst 2024) – zu finden in der Rubrik »Fokus«.

Online weiterlesen – mit interaktiven Grafiken, methodischem Annex und mehr Analysen: **www.machine-vs-rage.net.**

MACHINE AGAINST THE RAGE

KI-Bilder im rechtlichen Kontext

KI-generierte Bilder können, wenn sie auf sehr großen Online-Plattformen (VLOPs) oder Suchmaschinen (VLOSEs) verbreitet werden, »systemische Risiken« durch das Design oder die Funktionsweise einer Plattform nach dem Digital Services Act (DSA) darstellen. Darunter fallen zum Beispiel Äußerungen mit »tatsächliche[n] oder absehbar nachteilig[en] Auswirkungen auf die gesellschaftliche Debatte und auf Wahlprozesse «. Risikominderungsmaßnahmen und eine klare Verantwortungspflicht von Telegram sind Teil der Lösung, aber ohne mehr Druck der EU-Kommission, den Messenger-Dienst als VLOP einzustufen, reine Zukunftsmusik. Solange gilt für solche unangenehmen, aber nicht offensichtlich rechtswidrigen Inhalte der Grundsatz »lawful but awful« - und der Schutz von Art. 5 Abs. 1 GG.

Viele Bildgeneratoren nutzen schon eingebaute unsichtbare Wasserzeichen oder kryptografisch signierte Metadaten, um das Täuschungsrisiko zu minimieren. Für das bloße Auge sind diese zusätzlichen und Sicherheitsmaßnahmen nicht erkennbar. Eine klar sichtbare Kennzeichnung von KI-generierten und -gestützten Bildern, so wie sie im AI Act gefordert ist, sagt nichts darüber aus, ob ein Inhalt irreführend oder manipulativ ist, sondern nur, dass er von einem KI-Bildgenerator erstellt wurde. Das zeigt auch, wie schwierig es im Moment noch ist, generative KI und deren Urheber*innen auf externen Plattformen adäquat zu regulieren, denn die wenigsten extremistischen Akteur*innen halten sich an Kennzeichnungen.

Fest steht: Extremistisches und strafrechtlich relevantes Material findet sich im Datensatz vergleichsweise selten. Was es dafür gibt, ist ein Trend hin zur Bestätigung der eigenen Weltsicht, unabhängig vom Wahrheitsgehalt.

Über die BAG

Um Maßnahmen gegen digitalen Hass proaktiv und wirkungsvoll gestalten zu können, unterstützt die Bundesarbeitsgemeinschaft »Gegen Hass im Netz« die Zivilgesellschaft mit wissenschaftlichen Instrumenten. Sie steht für vernetztes Handeln und hat eine evidenzbasierte Praxis gegen digitalen Hass zum Ziel. Zu diesem Zweck unterhält sie eine hauseigene Forschungsstelle und vereinigt Akteure aus der Praxis in einem zivilgesellschaftlichen Forum. Die Wissenschaft liefert hierbei der Zivilgesellschaft Reflexionswissen – und andersrum fließt Praxiswissen in die Forschungsstelle ein.

Über die Forschungsstelle

Die Instrumente, um digitalen Hass besser zu verstehen, liefert uns die Digitalisierung selbst. In der Forschungsstelle der BAG kommen längjährige Erfahrung in der Extremismusforschung mit daten- und netzwerkanalytischer Expertise zusammen. So entsteht ein Monitoringsystem, das Trends in den Netzwerken des Hasses direkt erkennbar und über lange Sicht besser einschätzbar macht. Begleitet wird die Arbeit von externen Wissenschaftler*innen, die die Forschung mitentwickeln und die Methoden evaluieren. Zehn Expert*innen aus verschiedenen Disziplinen stehen hierbei beratend zur Seite.

Über den Trendreport

Die Ergebnisse und Analysen des Monitorings werden alle drei Monate in einem digitalen Trendreport veröffentlicht. Machine Against the Rage, so der Name dieses Online-Magazins, ist damit das zentrale Organ der Forschungsstelle. Es fungiert zum einen als Trendbarometer, mit dem wichtige Verschiebungen und Online-Aktivitäten in rechtsextremen und anderen demokratiefeindlichen Diskursen frühzeitig identifiziert werden. Zum anderen werden darin kritische Veränderungen der Meinungsentwicklung in relevanten Online-Milieus dokumentiert und analytisch eingeordnet.

Träger der BAG »Gegen Hass im Netz« ist
Das NETTZ | Vernetzungsstelle gegen Hate Speech

DAS NETTZ

www.das-nettz.de

Gefördert vom

Bundesministerium für Familie, Senioren, Frauen und Jugend

im Rahmen des Bundesprogramms

Demokratie leben!

Die Veröffentlichung stellt keine Meinungsäußerung des BMFSFJ, des BAFzA oder anderer Förderpartner*innen dar. Für inhaltliche Aussagen und Meinungsäußerungen tragen die Publizierenden dieser Veröffentlichung die Verantwortung.

https://machine-vs-rage.net

Bundesarbeitsgemeinschaft »Gegen Hass im Netz«
Redaktion: Lena-Maria Böswald, Christian Donner, Maik Fielitz

Das NETTZ gGmbH
c/o betterplace Umspannwerk GmbH
Paul-Lincke-Ufer 21, 10999 Berlin
E-Mail: info@bag-gegen-hass.net

www.bag-gegen-hass.net | www.das-nettz.de

Geschäftsführung: Nadine Brömme, Hanna Gleiß
Registergericht: Amtsgericht Berlin Charlottenburg, HRB 242638 B
Geschäftssitz: Berlin