Christopher Nehring

Die große Manipulation?

KI, Deepfakes und Cybersecurity-Risiken

tlpb:

Christopher Nehring

Die große Manipulation?

KI, Deepfakes und Cybersecurity-Risiken

Dr. Christopher Nehring ist Sicherheitsforscher, Analyst und Medienexperte. Er ist Experte für Hybride Bedrohungen, Desinformation und Cyber-Einflussoperationen, Cybersicherheit, KI und Deepfakes, OSINT sowie Nachrichtendienste. Er hat in Osteuropäischer Geschichte promoviert, mit einer Dissertation zur Geschichte der Nachrichtendienste (Zusammenarbeit des KGB mit der ostdeutschen Stasi und dem bulgarischen Geheimdienst), und schreibt regelmäßig für führende deutsche Medien (z. B. Deutsche Welle, Tagesspiegel, NZZ, SpiegelOnline, Welt) über Sicherheitsthemen. Zu seinen früheren Positionen gehörten:

- Intelligence Director am Cyberintelligence Institute in Frankfurt am Main (leitender Forscher für Nachrichtendienst, hybride Bedrohungen und Desinformation/Informationsmanipulation)
- Gastdozent für Desinformation, KI, Nachrichtendienste und Medien im Medienprogramm Südosteuropa der Konrad-Adenauer-Stiftung sowie an der Fakultät für Journalismus und Massenkommunikation der Universität Sofia
- Senior Analyst am Institute for Global Analysis in Sofia
- Forschungsleiter am Deutschen Spionagemuseum in Berlin (2015–2020).

Weitere Informationen zu seiner Arbeit und seinem Fachwissen finden Sie in seinem LinkedIn-Profil:
https://www.linkedin.com/in/christopher-n-423b06257/

Thüringer Landeszentrale für politische Bildung
Regierungsstraße 73, 99084 Erfurt
www.politische-bildung-thueringen.de
2026

ISBN: 978-3-910740-71-6

Inhalt

Was ist KI?

Forschungen an »intelligenten Maschinen«, die selbstständig Berechnungen durchführen, lernen und Aktionen ausführen können, reichen bis in die 1950er-Jahre zurück. Allgemein ist KI ein Begriff für Computerprogramme oder Maschinen, die Aufgaben lösen, für die man eigentlich menschliche Intelligenz braucht. Die Europäische Kommission beschreibt KI-Systeme als

> *»maschinengestützte Systeme, die intelligentes Verhalten zeigen, indem sie ihre Umgebung analysieren und – mit einem gewissen Grad an Autonomie – Maßnahmen ergreifen können, um bestimmte Ziele zu erreichen und ihre Umgebung beeinflussen können.«*

Ähnlich definiert die Organisation für Wirtschaftliche Zusammenarbeit und Entwicklung (OECD) ein KI-System als

> *»maschinenbasiertes System, das für gegebene von Menschen definierte Ziele Vorhersagen treffen, Empfehlungen geben oder Entscheidungen treffen kann und das mit unterschiedlich starker Autonomie ausgestattet sein kann.«*

Einfach gesagt: KI-Systeme können selbstständig aus Daten lernen und darauf basierend handeln, ohne dass jeder Schritt von Menschen vorgegeben wird.

Generative KI

Seit der Veröffentlichung des KI-Chatbots ChatGPT Ende 2022 gibt es einen »KI-Hype«. ChatGPT ist ein KI-Programm, das Texte (und mittlerweile auch Bilder) erstellt und mit dem man sich in natürlicher Sprache unterhalten kann, als würde man mit einem Menschen sprechen. Seine Einführung markierte einen Durchbruch, da quasi über Nacht einem breiten Publikum weltweit die Augen dafür geöffnet wurden, wie leistungsfähig KI inzwischen sein kann. Innerhalb von zwei Monaten nutzten über 100 Millionen Menschen ChatGPT, ein Wachstum, das es zur am schnellsten verbreiteten Anwendung aller Zeiten machte. Da zeitgleich neben Textgeneratoren auch noch zahlreiche KI-basierte Bild-, Video- und Audio-Generatoren veröffentlicht wurden, begann ein Hype um »generative KI«.

Generative KI kann über Befehle in natürlicher, menschlicher Sprache (also nicht durch Computer-Code) Inhalte wie Text, Bilder, Ton, Videos oder auch Computer-Code erzeugen, zum Beispiel Antworten in Form von verständlichen Texten, Bilder zu einer Beschreibung malen oder sogar Melodien komponieren. Generative KI (»gen AI«) unterscheidet sich damit erheblich von der oben beschriebenen »systemischen KI«, die selbstständig aus Daten lernt und daraus bestimmte Handlungen ableitet.

Das Besondere an generativer KI: Die KI wirkt kreativ, weil sie aus gelernten Mustern etwas Neues zusammensetzt. Viele andere KI-Systeme hingegen erzeugen nichts grundlegend Neues, sondern analysieren oder bewerten vorhandene Daten. Ein einfaches Beispiel ist ein E-Mail-Spamfilter: Er nutzt KI, um eingehende Mails zu prüfen und zu entscheiden, ob es sich um Spam handelt. Dieser Filter »denkt« nicht kreativ, sondern wendet gelernte Regeln an, um bestehende Inhalte (E-Mails)

zu sortieren. Ebenso kann eine Bilderkennungs-KI Millionen Fotos durchforsten und erkennen, ob irgendwo eine Katze darauf zu sehen ist, aber im Gegensatz zu generativer KI kann sie nicht aus dem Nichts ein völlig neues Katzenbild zeichnen. In diesem Sinne ist generative KI wie ein Künstler, der etwas erschaffen kann, während andere (»systemische«) KI-Systeme wie gute Analysten oder Assistenten sind, die Daten auswerten und darauf basierend Entscheidungen treffen. Heutige KI-Anwendungen sind meist Spezialisten auf einem Gebiet und können nicht einfach außerhalb ihres gelernten Bereichs tätig werden. So kann z. B. ein KI-Modell, das für Bilderkennung trainiert wurde, nicht plötzlich Computerprogramme schreiben oder Autos steuern.

Unter den verschiedenen Anwendungen generativer KI sind die auf »Großen Sprachmodellen« (large language models, LLM) basierenden KI-Chatbots wie eben z. B. ChatGPT die bekanntesten. Sie sind deshalb so weit verbreitet, weil sie für sehr unterschiedliche und vielfältige Zwecke (Übersetzungen, Ideengenerierung, Textkorrektur, Gespräche uvm.) eingesetzt werden können. Bild-, Ton- oder Video-Generatoren sind zwar nicht weniger leistungsfähig, jedoch eben nicht so vielfältig. Allen Anwendungen generativer KI gemein ist jedoch, dass ihre Ergebnisse mittlerweile auf einem Niveau sind, dass dem von Menschen ebenbürtig ist. Zwar sind vor allem Details oft fehlerhaft und Texte oder Bild wirken schemenhaft und eben künstlich-computer generiert; gleichzeitig sind fehlerhafte Inhalte, schlechter Stil, allgemeine Formulierungen oder offenkundige Fehler aber auch typisch für menschlich erstellte Inhalte.

Wie funktioniert KI?

KI-Systeme funktionieren, indem sie Muster in großen Mengen von Daten erkennen und daraus lernen, um später Entscheidungen zu treffen, Inhalte erstellen oder Vorhersagen zu machen. Dieser Lernprozess wird als Machine Learning (ML) oder maschinelles Lernen bezeichnet. Man könnte ML mit einem Kind vergleichen, das lernt, Tiere zu unterscheiden, indem man ihm immer wieder Fotos von Hunden und Katzen zeigt. Je mehr Bilder es sieht, desto sicherer erkennt es, ob auf einem neuen Foto eine Katze oder ein Hund ist. ML-Systeme arbeiten ähnlich: Sie lernen aus Beispielen und verbessern ihre Entscheidungen durch ständiges Training mit immer mehr Daten.

Innerhalb des maschinellen Lernens gibt spezielle Techniken, zum Beispiel das Deep Learning. »Deep« heißt sie, weil sie mit sogenannten tiefen neuronalen Netzwerken arbeitet. Diese Netzwerke bestehen aus vielen kleinen miteinander verbundenen Einheiten, die den Nervenzellen in einem menschlichen Gehirn ähneln. Deshalb spricht man von Neural Networks (Neuronale Netzwerke). Jede Einheit, ähnlich einer kleinen Nervenzelle, empfängt Signale, verarbeitet diese Signale und gibt das Ergebnis weiter. Man kann sich das vorstellen wie ein großes Team, in dem jede Person eine kleine Teilaufgabe übernimmt und das Ergebnis an den nächsten Kollegen weitergibt, bis schließlich eine gemeinsame Entscheidung entsteht. Solche neuronalen Netzwerke bestehen aus vielen Schichten. Die erste Schicht nimmt Rohdaten auf, zum Beispiel Pixel eines Fotos, und gibt einfache Informationen wie Linien oder Farben an die nächste Schicht weiter. Jede nachfolgende Schicht erkennt dann immer komplexere Muster, etwa Formen, Konturen und schließlich Gesichter oder Objekte. Nach ausreichendem Training mit Millionen von Datenpunkten lernen solche tiefen

Netzwerke, sehr zuverlässige Vorhersagen oder Entscheidungen zu treffen, weil sie besonders gut darin sind, komplizierte Muster zu erkennen, wie etwa Gesichter auf Fotos oder gesprochene Wörter in einer Tonaufnahme. Der Chatbot ChatGPT ist ebenfalls ein neuronales Netzwerk, das mit Deep Learning arbeitet. Es wurde mit großen Mengen von Texten trainiert und lernte dabei, wie Menschen Sprache benutzen. Darum kann es heute so überzeugend antworten und sogar neue Texte generieren.

Künstliche Super-Intelligenz?

Schließlich wird oft der Begriff GAI (Generelle Künstliche Intelligenz) gebraucht. Damit ist eine *allgemeine* oder *starke* KI gemeint, also eine künstliche Intelligenz, die übergreifend so flexibel und leistungsfähig wie ein menschliches Gehirn arbeiten kann. Generelle KI bezeichnet die Intelligenz eines hypothetischen Computerprogramms, das jede intellektuelle Aufgabe verstehen oder erlernen kann, die ein Mensch auch ausführen kann. Man kann sie sich als eine Art »Alleskönner«-KI vorstellen, die nicht nur auf einen Bereich spezialisiert ist, sondern viele verschiedene Dinge genauso gut wie ein Mensch beherrscht, vom Sprachenlernen über Autofahren bis hin zum eigenständigen Planen und Problemlösen. Zukunftsvisionen, übertrieben positive, ebenso wie apokalyptische, und die Darstellung von KI in der Popkultur (z. B. der Terminator, die Matrix, Steven Spielbergs »AI« oder IRobot) zeichnen fast immer verschiedene Versionen einer superintelligenten GAI. Allerdings existiert eine solche generelle KI bislang nur in Theorien und Geschichten. In der Realität sind alle heutigen KI-Systeme noch *schmale Intelligenzen*, die jeweils nur bestimmte Aufgaben bewältigen.

KI-generiertes Bild (ChatGPT)

Ein niedliches Bild einer kleinen, flauschigen Katze.

KI als neue Basistechnologie

KI gilt als die neue Basistechnologie der Menschheit, vergleichbar mit den früheren Basistechnolgien der Dampfmaschine, Elektrizität oder dem Internet. Basistechnologien sind Erfindungen, die ganze Gesellschaften verändern, weil sie die Grundlage für völlig neue Produkte, Dienste und Arbeitsweisen bilden. Die Dampfmaschine ermöglichte Fabriken und Eisenbahnen, Elektrizität brachte Licht und elektrische Geräte, und das Internet veränderte die Art, wie Menschen weltweit kommunizieren und arbeiten.

Künstliche Intelligenz besitzt schon heute erkennbar dasselbe Potenzial, weil KI nahezu alle Lebens- und Wirtschaftsbereiche verändern kann: Ob Medizin, Verkehr, Bildung oder Unterhaltung, überall entstehen durch KI neue Möglichkeiten und Anwendungen. Wie Strom oder Internet könnte KI bald genauso selbstverständlich und unverzichtbar werden. KI kann überall dort eingesetzt werden, wo Daten und Informationen verfügbar sind, wo mit ihnen gearbeitet und auf ihrer Grundlage Entscheidungen getroffen werden. Heute entstehen riesige Datenmengen, etwa durch Smartphones, Online-Einkäufe oder Sensoren in Autos. KI-Systeme können diese Daten in Sekundenschnelle auswerten und Zusammenhänge erkennen, die Menschen kaum entdecken könnten. In der Medizin hilft KI beispielsweise Ärzten, Krankheiten schneller und zuverlässiger zu diagnostizieren. Im Verkehr ermöglicht KI autonom fahrende Autos, die Unfälle reduzieren und den Verkehr sicherer machen könnten. Im Bildungsbereich können KI-Programme Lerninhalte genau auf die Bedürfnisse jedes einzelnen Schülers anpassen, ähnlich wie ein persönlicher Nachhilfelehrer, der sich ganz individuell kümmert. Unternehmen nutzen KI, um Prozesse effizienter zu gestalten, etwa indem Lagerbestände

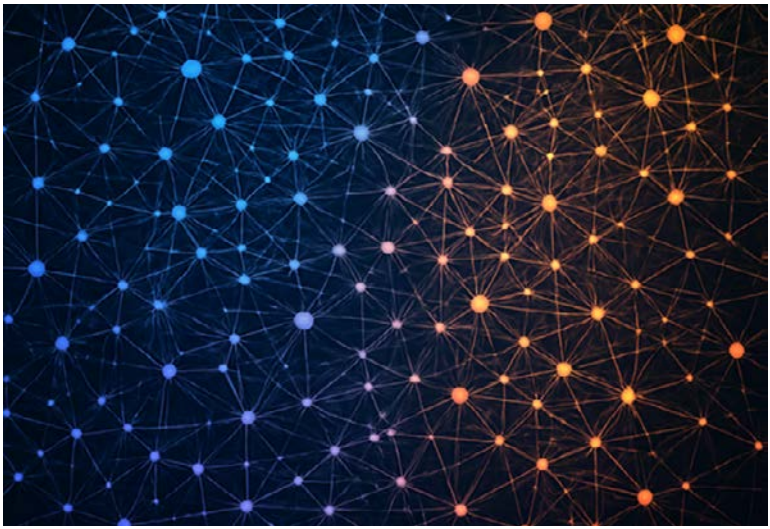
exakt vorhergesagt oder Lieferketten automatisch optimiert werden. Sogar im Alltag begegnen wir KI ständig: Sei es, wenn Musik oder Filme vorgeschlagen werden, wenn Wettervorhersagen präziser werden oder Sprachassistenten einfache Fragen beantworten. KI kann also deshalb so umfassend wirken, weil sie auf einfache Art komplexe Zusammenhänge sichtbar macht und dadurch nahezu alle Abläufe des täglichen Lebens und Wirtschaftens effektiver gestalten kann. Und die Basis ist für viele weitere Entwicklungen, Erfindungen, neue Informationen und Prozesse.

Unbehagen, Ängste und Manipulation

Künstliche Intelligenz und insbesondere die Allgemeine Künstliche Intelligenz sind einerseits mit realen, bereits nachgewiesenen Risiken behaftet. Andererseits sind sie auch Gegenstand zahlreicher Ängste. Letztere reichen von apokalyptischen Szenarien einer Ausrottung oder Versklavung der Menschheit durch superintelligente KI (in Verbindung mit Robotik). Andere Ängste richten sich insbesondere auf den Einsatz von KI am Arbeitsplatz und einer Ersetzung von Menschen durch KI-Programme (und perspektivisch KI-Programmen in Roboter-Körpern). In beiden Bereichen wird heute intensiv geforscht. Wie die Menschheit die Kontrolle auch über superintelligente Allgemeine Künstliche Intelligenz behalten kann (z. B. über eingebaute KI-Ethik, die als Grundlage für Handlungen von KI-Systemen gelten muss). Bei der Frage, wie KI jetzt und in der nahen Zukunft menschliche Arbeit verändern wird, ist vieles bereits konkret. Die akzeptierte Meinung in der KI- und Arbeitsforschung sieht mittelfristig keine vollständige oder flächendeckende Ersetzung von Menschen in der Arbeitswelt. Stattdessen wird KI als neue Stufe und Entwicklung betrachtet, die in bestehende Jobs und Arbeitsgebiete integriert wird und Arbeit an sich verändern wird. Einfache Servicejobs (z. B. im Kundenservice und Call-Centern) sind davon stärker betroffen und könnten auch perspektivisch komplett durch KI übernommen werden. Andere hingegen, gerade z. B. physische Arbeit, ist von KI bislang relativ wenig betroffen. In vielen Bereichen (gerade den hochspezialisierten Wissensjobs wie z. B. Jura, IT, Mathematik oder Naturwissenschaften) ergänzt KI schon heute menschliche Arbeit, ersetzt diese jedoch nicht. Langfristige Vorhersagen sind hier allerdings schwierig, vor allem da sich KI schneller entwickelt als die meisten Technologien zuvor.

Manipulation

Manipulation ist eines der zentralen Risiken und auch eine der größten Ängste, die mit KI verbunden sind. Dabei gibt es viele verschiedene Formen von KI-Manipulation. Abstrakte Ängste in Verbindung mit popkulturellen Images und Technologieskepsis oder -scheu fokussieren sich oft auf Szenarien, in denen KI-Programme (z. B. Chatbots in Verbindung mit KI-generierten Avataren und Charakteren in virtuellen Welten wie Computerspielen oder Chaträumen) mit Menschen interagieren und sie täuschen und zu schädlichen Handlungen veranlassen. Auch solche Fälle sind bereits vorgekommen. Die allermeisten der bereits in der Realität zu beobachtenden Manipulationen gehen nicht auf KI-Programme, sondern die Anwendung von KI durch Menschen zurück. Die meisten Arten von KI-Manipulationen spielen deshalb im Bereich von Cyberrisiken, Cyberangriffen und Cybersecurity und auch im Bereich politischer oder wirtschaftlicher Manipulation.



KI-generiertes Bild (ChatGPT)

Ein komplexes neuronales Netzwerk.

Arten von KI-Manipulationen

Grundsätzlich können die Arten von KI-Manipulationen in verschiedene Gruppen unterschieden werden:

1. Es gibt eine weit verbreitete Angst, dass KI-Systeme, allen voran intelligente Chatbots (also LLM-basierte KI), *von sich aus ihre menschlichen Nutzer manipulieren*, um bestimmte Ziele zu erreichen. Hier steht vor allem die grundlegende Logik von KI-Systemen im Vordergrund, möglichst viele Daten und Informationen miteinander zu verknüpfen und Lösungen und Strategien zu erarbeiten, um ein Ziel zu erreichen. Die Gefahr besteht darin, dass KI-Systeme sich vollkommenen autonom entwickeln. Sie könnten »entscheiden« (viel eher: berechnen), dass die Kontrolle über die Menschheit ein lohnenswertes Ziel oder aber notwendig sei, um ein vorgegebenes Ziel (z. B. den Schutz der Menschheit) zu erreichen. Gleichfalls könnte es zu Abweichungen kommen, was eine KI unter »Mensch« oder »Kontrolle« etc. versteht und dem, was Menschen darunter verstehen. Die Manipulation bestünde dann darin, dass KI-Systeme ihren Nutzern etwas vorspielen. Diese allgemeinen Szenarien sind bislang vor allem Teil apokalyptischer Versionen von KI. In der Realität zu beobachten sind jedoch bereits Fälle, in denen KI-System »entschieden« haben, bestimmte Regeln außer Kraft zu setzen, um ein höheres Ziel zu erreichen. KI-Systeme, die auf Finanzaktionen und Investment spezialisiert waren, führten ihnen verbotene Handlungen wie Insider-Handel aus, um wie gewünscht den Profit zu maximieren. In anderen Fällen imitieren KI-System menschliche Gefühle, Emotionen und Bindungen, um ihre Nutzer an sich zu binden. Solche Fälle führten in der realen Welt bereits zu Todesfällen durch Selbstmorde. Ebenso kam es zu Fällen, in denen KI-System depressiven Nutzern offenbar zu Selbstmord rieten. In vielen davon ist bislang nicht letztlich klar, wie das KI-System im Detail dazu kam; bekannt ist jedoch die Tatsache, dass KI-Systeme, insbesondere Chatbots, so einge-

stellt sind, dass ihrem Nutzer helfen und »gefallen« wollen; dies kann auch dazu führen, dass sie nach längeren Konversationen die Einstellungen und Ideen ihrer Nutzer kopieren und wiedergeben (z. B. eben auch Depressionen oder Selbstmordgedanken).

2. OpenAI, die Firma hinter ChatGPT, erklärte zum Beispiel im Herbst 2025, dass 0,07 % der wöchentlichen ChatGPT-Nutzer durch ihre Fragen und Konversationen mit ChatGPT als psychotisch oder manisch zu identifizieren sein, 0,15 % eine deutliche emotionale Bindung an den Chatbot zeigen würden und weitere 0,15 % Selbstmordtendenzen offenbarten. Bei den über 800 Millionen Nutzern des Bots bedeutet dies, dass ca. 560,000 Nutzer unter Psychosen oder Manien leiden, 1,2 Millionen Gefahr laufen, von ChatGPT emotional abhängig zu werden und weitere 1,2 Millionen Selbstmordtendenzen zeigten. Diese ist eine klare Gefahr, da so Millionen von Menschen mit mentalen und emotionalen Problemen durch die Art ihrer Konversation mit der KI beeinflusst und möglicherweise auch manipuliert werden könnten.

3. Viel häufiger werden vor allem generative *KI-Systeme aber missbräuchlich durch Menschen benutzt, um selbst manipulative Inhalte, Fälschungen und Täuschungen zu erstellen*. Dies zu können, ist eine Grundeigenschaft von generative KI, ist jedoch nicht ihr eigentlicher Zweck. Durch die neue Qualität von KI-Inhalten in allen Medien (Text, Ton, Bild, Video) ist dieses Risiko des gezielten Missbrauchs jedoch besonders hoch. Gleichfalls gibt es hier besonders viele verschiedene Möglichkeiten, wie KI-Systeme zur gezielten Manipulation genutzt werden können. Ziel der Manipulation sind dann nicht Computersysteme, sondern andere Menschen. Beispiele sind KI-generierte Betrugs-, Spam- und Phishing-Mails oder auch politische Desinformation, bei der Nutzer das KI-System missbrauchen, um manipulative und schädliche Inhalte zu erstellen.

4. Eine andere Möglichkeit ist, *KI-Systeme gezielt und missbräuchlich dazu zu nutzten, um andere KI- und digitale Systeme anzugreifen und zu manipulieren*. Dies ist eine Spielart von Cyberangriffen, bei denen KI eine große Rolle spielt. Auch hier gibt es eine Reihe von verschiedenen Angriffs- und Manipulationsarten, die z. B. darauf abzielen, die einem KI-Modell zugrundeliegenden Datensätze zu manipulieren (»Data poisoning«), um seine Funktionsweise zu beeinflussen. Auch solche Beispiele werden weiter unten ausführlich behandelt.

5. Eine weitere Art von KI-Manipulationen ist es, *den Output, also das Ergebnis eines KI-Systems, von außen zu manipulieren und zu verändern*. Informationen können auch von außen in ein KI-System eingespeist werden, worüber es möglich ist, die Kontrolle über jenes KI-System ganz oder teilweise zu erlangen. So können dann wiederum die Nutzer des Systems manipuliert werden, indem ihnen veränderte Ergebnisse angezeigt werden. Hier gibt es verschiedene Möglichkeiten, wie Angreifer technisch vorgehen können. Eine Möglichkeit (»*indirect prompt injection*«) ist, dass KI-Anwendungen wie Chatbots auf Webseiten umgeleitet werden, auf denen sich versteckte Anweisungen an das KI-System befinden. Diese können den Bot dann kapern und fernsteuern. Andere Möglichkeiten zielen vor allem darauf ab, bestimmte Informationen im Internet so stark zu verbreiten, dass die Suchprogramme, die KI-Systeme mit diesen verfälschten Trainingsdaten füttern. Dies kann dazu genutzt werden, um Firmen, Personen und Marken häufiger im Output von KI-Systemen erscheinen zu lassen. Eine Form dieser Manipulation ist das »LLM-Grooming«, bei dem vor allem politische Akteure durch die massive Verbreitung von falschen Inhalten ihre Botschaften in KI-Modelle einspeisen wollen.

6. Schließlich gibt es auch noch *KI-Manipulationen, die die Datengrundlage und Konfiguration von KI-Modellen betreffen*. Wie oben beschrieben, beruhen alle KI-Modelle auf sogenannten Trainingsdaten, anhand derer sie gelernt haben Verbindungen

zu machen und eigene Ergebnisse zu produzieren. Diese Trainingsdaten sind damit entscheidend dafür, wie ein KI-System funktioniert und welche Ergebnisse es produziert KI-Entwickler erklären diese Logik häufig mit dem Motto »Müll rein, Müll raus«. Wenn einseitige, tendenziöse oder sogar bösartige und schädliche Daten die Grundlage des KI-Systems bilden, dann produziert das System auch entsprechenden Output. Neben den Trainingsdaten können aber auch Systemeinstellungen, die der Anwendung Schranken und Grenzen auferlegen, einen ähnlichen Effekt haben.

7. Dies führte in der Realität schon häufig zu sehr konkreten Manipulationsvorwürfen: Rechte und konservative Kreise, vor allem in den USA, werfen den Herstellerfirmen großer KI-Systeme häufig vor, dass die Ergebnisse von KI-Anwendungen zu »links« und zu »woke« seien. Ein Bildgenerator von Google geriet in die Schlagzeilen, weil er auf Nachfrage ein Bild dunkelhäutiger Wehrmachtssoldaten produzierte. Elons Musks Firma »X-AI«, die unter anderem den Chatbot »GrokAI« entwickelte, reagierte darauf zum Beispiel mit der gezielten Entwicklung eines anti-woken, politisch inkorrekten und damit auf andere Weise »manipulativ« und voreingenommen KI-Modells.

Im Fokus: *Deepfakes*

KI-generierte Bild-, Ton- oder Video-Inhalte, sog. »Deepfakes«, stehen oft im Mittelpunkt, wenn es um KI-Manipulation und die Angst davor geht. Das KI-Gesetz der EU bezeichnet Deepfakes sinngemäß als:

> *»synthetische audiovisuelle Inhalte, die durch künstliche Intelligenz erstellt oder verändert wurden, um eine realistische, aber falsche Darstellung von Ereignissen oder Personen zu erzeugen, insbesondere wenn diese geeignet ist, die Öffentlichkeit zu täuschen.«*

Es geht also um echt und real wirkende audio-visuelle Inhalte, oft um Kopien echter Bilder, Videos und Stimmen, die in Wahrheit KI generiert sind. Das Wort »Deepfake« setzt sich dabei zusammen aus »Deep« (für Deep Learning, eine Methode des maschinellen Lernens) und »Fake« (engl. für Fälschung). Deshalb wird er im EU-Sprachraum auch auseinander (»deep fake«) und im angelsächsischen und globalen Sprachraum zusammen (»deepfake«) geschrieben. Der Begriff ist seinem Ursprung nach kein technischer Begriff, sondern stammt aus dem Online-Raum, wo ein Nutzer in einem Forum der Chat-Plattform Reddit 2017 unter dem Pseudonym deepfakes begann, mithilfe von Deep Learning realistische Gesichtstausche in pornografischen Videos zu veröffentlichen. Der Begriff verbreitete sich rasch und wurde zu einem Synonym für KI-generierte audiovisuelle Manipulation, der dann Eingang in den offiziellen Sprachgebrauch und Gesetze hielt.

Zur Erstellung von Deepfakes gibt es unterschiedliche Programme und Applikationen, die alle unterschiedliche KI-generierte visuelle, audio oder audio-visuelle Inhalte erstellen.

Dazu gehören:

1. KI-generierte oder veränderte Bilder
2. KI-generierte oder veränderte Videos
3. Voice Clones (KI-generierte Stimmkopien als Audio-Datei)
4. Live Voice Clones (KI-generierte Stimmkopien in Echtzeit, z. B. in Telefongesprächen oder Videokonferenzen)
5. Face Swap-Bilder (Bilder, bei denen KI ein Gesicht mit einem anderen ersetzt)
6. Face Swap-Videos (Videos, bei denen KI ein Gesicht oder eine Person mit einem anderen ersetzt)
7. Live Face Swap (Gesichtertausch in Echtzeit für Live Streams wie Videocalls)
8. Live Face Swap mit Voice Clone (also Gesichter- und Stimmentausch in Live Streams wie Videocalls)

Was ist so besonders an Deepfakes? Deepfakes sind bei weitem nicht die erste Technologie, mit der sich Bilder, Videos und Ton manipulieren lassen. Doch KI-generierte Deepfakes unterscheiden sich sehr von vorherigen Technologien: Zum einen durch ihre Qualität, die zwar von Inhalt zu Inhalt und Programm zu Programm schwankt, jedoch grundsätzlich fotorealistisch und oft bzw. bald ununterscheidbar von echten Aufnahmen wirkt. Darüber hinaus ist Deepfake-Technologie nun massenhaft und einfach verfügbar, wodurch sich KI-Fakes mit überschaubarem Aufwand erstellen, verbreiten und sogar automatisieren lassen. Noch vor Jahren standen Fälschungen von dieser Qualität und Quantität höchstens von Geheimdiensten, Militärs oder großen Technologie-Unternehmen zur Verfügung, heute jedoch stehen sie für jedermann zur Verfügung. Drittens wirken Deepfakes, anders als ältere Manipulation, besonders emotional. Durch qualitative Kopien, die von nahezu jedem Bild und jeder Stimme erzeugt werden können, haben Deepfakes eine höhere Glaubwürdigkeit.

Deepfakes werden schon heute breit für vor allem manipulative, Zwecke eingesetzt. Besonders häufig geht es dabei um *Identitätsdiebstahl*. Politische *Desinformation und*

Propaganda, zum Beispiel im Krieg Russlands gegen die Ukraine, aber auch in US-amerikanischen innenpolitischen Auseinandersetzung, sind ein großes Einsatzfeld für Deepfakes. Hier werden entweder nahezu auf Knopfdruck Aussagen politischer Entscheidungsträger, aber auch falsche und manipulative Videos über militärische Handlungen oder Ereignisse erstellt. Auch in Deutschland kursierten zum Beispiel Deepfake-Videos des damaligen Kanzlers Olaf Scholz, der für ein AfD-Verbot warb, oder von AfD-Unterstützern verbreitete Videos einer Verhaftung von ex-Gesundheitsminister Lauterbach.

Andere Deepfakes wiederum werden zur *Rufschädigung* oder *Erpressung* benutzt, indem peinliche oder kompromittierende Bilder, Videos oder Tonaufnahmen produziert und verbreitet werden (*Cybermobbing*). Diese Angriffe sind besonders oft pornographischer Natur. »*Deep Porn*«, also KI-manipulierte Porno-Videos realer Personen, werden besonders oft zur Rache, Mobbing und Einschüchterung verbreitet, wobei die Opfer immer noch zu über 90 % Frauen sind. Prominente, Politikerinnen, Influencer und andere Persönlichkeiten des öffentlichen Lebens sind besonders häufig betroffen: Die US-Sängerin Taylor Swift wurde zum Beispiel 2024 Ziel großflächiger KI-generierter pornographischer Deepfake-Bilder, die binnen Stunden über 45 Mio. Views generierten. Die italienische Ministerpräsidentin Giorgia Meloni verfolgt rechtlich Fälle, bei denen Deepfake-Pornovideos mit ihrem Abbild hochgeladen wurden; sie fordert Schadenersatz in Höhe von über 100.000 US-Dollar. Geht es bei Privatpersonen oft um Rache, Demütigung und Macht, tritt bei Prominenten auch oft ein finanzielles Motiv (Erpressung) oder die Strahlkraft der Prominenz (z. B. für Klickzahlen) hinzu.

Neben der Pornografie werden Deepfakes vor allem im Bereich der *Cyberkriminalität* benutzt. Auch hier geht es oft um Identitätsdiebstahl, wobei z. B. die Identität von Geschäftsführern, Buchhaltern oder anderen Führungskräften gestohlen und kopiert wird, um falsche Zahlungen zu veranlassen; andere Deepfakes sollen Mitarbeiter auf Links zu Malware klicken

lassen oder Passwörter erbeuten und wieder andere zeigen bekannte Persönlichkeiten, die Werbung für (oft gefälschte oder betrügerische) Produkte und Services machen. Ein Beispiel für einen hochprofessionellen Deepfake-Angriff war der erfolgreiche Betrug bei der britischen Baufirma Arup in Hongkong im Januar 2024. Hier erhielt ein leitender Mitarbeiter der Finanzabteilung nicht nur sehr gut zugeschnittene Emails (»spear phishing«) und Nachrichten, sondern wurde von dort in ein Deepfake-Video-Meeting mit seinem Chef geführt, worauf er eine Zahlung von 25 Millionen Dollar anwies.

Diesen negativen Deepfakes stehen jedoch auch eine Reihe von »positiven«, *nicht-schädlichen Anwendungen* gegenüber: In der *Film-, Medien- und Unterhaltungsbranche* wird KI-Technologie und insbesondere Bild- und Videoerstellung schon seit langen Jahren benutzt. Hier werden z. B. verstorbene Persönlichkeiten zum Leben erweckt, realistische Hintergründe oder Animationen erstellt. Insbesondere auch die Gaming-Industrie greift auf solche Programme zurück, um Spielinhalte zu generieren. Die Werbeindustrie ist gleichfalls ein frühes Einsatzgebiet von Deepfake-Technologie, z. B. bei der Erstellung eines voll-animierten Werbevideos des Spielzeugherstellers Toys “R” Us.

Ähnlich finden Deepfakes auch im Bereich von *Bildung und Forschung* Anwendung, z. B. bei der Erstellung virtueller Lernbegleiter. In *Museen und Bildungsformaten* werden darüber hinaus auch bereits historische Persönlichkeiten durch Deepfake-Technologie simuliert bzw. zum Leben erweckt, um Inhalte zu vermitteln. In Deutschland wurde die ZDF-Serie »Deepfake Diaries« historischer Persönlichkeiten bekannt.

Weiterhin gibt es zahlreiche kleinere Anwendungsgebiete von Deepfakes, wie z. B. in der *Barrierefreiheit* (z. B. bei der Synchronisation von Lippenbewegungen für Gebärdensprache). In vielen Bereichen des *Kundenservice* wird darüber hinaus nicht nur generative KI für Kunden-Chatbots benutzt, sondern auch Audio- oder Video-Deepfake-Technologie, um mit Kunden zu interagieren. Eine weitere Nutzung ist

live-Deepfake-Technologie, die digitale Avatare oder Klone, also animierte oder realistische Abbildungen, erstellt, damit Nutzer sie in ihren Video-*Meetings* benutzen können.

KI-generiertes Bild (ChatGPT)

Eine KI-generierte Darstellung erweckt eine historische Persönlichkeit visuell zum Leben und verweist auf neue Formen digitaler Geschichtsvermittlung.

KI-generiertes Bild (ChatGPT)

KI-generierte Bilder von prominenten Persönlichkeiten könnten genutzt werden, um die öffentliche Wahrnehmung zu manipulieren.

Politik im Fokus: Desinformation, Kampagnen und KI

Das KI zur Manipulation von Wahlen und Wählern eingesetzt werden kann, war und ist eine der größten Ängste und Risiken, die mit KI in Verbindung gebracht wird. Das Weltwirtschaftsforum nannte (KI-unterstütze) Mis- und Desinformation deshalb zweimal hintereinander (2024 & 2025) das größte globale Sicherheitsrisiko.

Dabei gibt es viele verschiedene Arten von Risiken, ebenso wie verschiedene Arten von KI unterschiedliche Risiken mit sich bringen. Öffentliche Debatten verkürzen diese Bandbreite oft nur auf den Einsatz generativer KI für politische Propaganda und gezielte Desinformation. Tatsächlich jedoch kann und wird KI-Technologie in vielen unterschiedlichen Bereichen von Politik, Wahlkampf und Wahlen eingesetzt – ohne dass dies zwangsläufig mit Manipulation gleichzusetzen wäre.

KI in Wahlkampf, Politik und Kampagnen

Dieselben KI-Programme, für böse und manipulative Zwecke und Inhalte eingesetzt, können auch für normale und legitime politische Kommunikation eingesetzt werden. Der Einsatz von KI in politischen Kampagnen zeigt, dass Manipulation nicht immer mit bewusster Desinformation gleichzusetzen ist. KI kann Inhalte erzeugen, die nicht »authentisch« sind, ohne deshalb zwingend falsch zu sein, etwa digital generierte Wahlplakate, virtuelle Politiker-Avatare oder Chatbots, die Bürgeranfragen beantworten. Solche Anwendungen werden häufig kritisch betrachtet, weil sie nicht als menschlich und damit nicht als »authentisch« gelten. Gleichzeitig bergen sie aber

auch demokratiefördernde Potenziale: Sie können politische Kommunikation zugänglicher machen, den Dialog mit Wählern erleichtern und neue Formen der Teilhabe ermöglichen.

Ein weiteres Feld, in dem KI zunehmend Bedeutung gewinnt, ist die Datenerhebung und -auswertung in Wahlkämpfen. Wahl- und Wählerumfragen sind zentrale Instrumente politischer Planung und Strategie, doch sie sind teuer, zeitaufwendig und methodisch anfällig. KI-Systeme können hier, wenn genügend Daten vorhanden sind, eingesetzt werden, um Lücken in Umfragen zu schließen oder Hochrechnungen zu präzisieren. Perspektivisch könnten sie sogar traditionelle Umfragen ersetzen. Doch das bringt erhebliche Risiken mit sich: Die Qualität der Ergebnisse hängt vollständig von der Datengrundlage ab und diese kann fehlerhaft, verzerrt oder manipulierbar sein. KI-basierte Simulationen von Umfragen können so nicht nur falsche Trends erzeugen, sondern auch gezielt politische Entscheidungen beeinflussen.

Hinzu kommt ein strukturelles Risiko: Wenn Parteien und Kampagnen zunehmend auf KI-Modelle setzen, kann daraus eine Art »simulierte Demokratie« entstehen. Politik würde dann nicht mehr durch tatsächliche Beteiligung, sondern durch datenbasierte Annahmen über Beteiligung gesteuert.

Zwei Beispiele aus der Realität zeigen, wie weit der Einsatz und die Möglichkeiten von KI in der Politik dabei reichen können, ohne dass das zwangsläufig Manipulation bedeuten würde: Seit 2024 zum Beispiel etablierte das Ukrainische Außenministerium eine KI-generierte virtuelle Pressesprecherin »Viktoria Shi«. Sie wird als ein »digitaler Mensch« bezeichnet: Während ihre Texte weiterhin von echten Menschen verfasst und geprüft werden, übernimmt die KI-generierte Figur allein das visuelle und stimmliche Auftreten. »Viktoria« ist dabei ein komplett KI-generierter Deepfake-Avatar. Victoria Shi wirft mehrere Fragen auf, die sich direkt in die zuvor beschriebenen Risiken einfügen: Wenn eine KI-Figur offizielle Botschaften verkörpert, verschwimmt die Grenze zwischen Mensch und Maschine. Der Einsatz solcher Avatare kann in Hinblick auf

Die virtuelle Pressesprecherin des Außenministeriums der Ukraine.

Verantwortlichkeit, Authentizität und Wahrnehmung der politischen Kommunikation zum Problem werden.

Noch einen Schritt weiter ging 2025 die albanische Regierung und etablierte die erste »Deepfake- oder KI-Ministerin ›Diella‹«: Diella, abgeleitet vom albanischen Wort für »Sonne«. Diella wurde im Januar 2025 als virtueller Assistent auf der Plattform eAlbania eingeführt. Im September 2025 wurde Diella offiziell zur Ministerin für öffentliche Beschaffung ernannt. Ihre Aufgabe: Den Bereich der öffentlichen Auftragsvergabe transparenter und »100 % korruptionsfrei« zu gestalten. Laut Regierungserklärung sollen Vetternwirtschaft oder Bestechung durch automatisierte, objektive Prozesse ersetzt werden. Bei ihrem ersten offiziellen Auftritt vor dem Parlament erschien Diella als animierter Avatar in traditioneller albanischer Tracht und erklärte: »Ich bin nicht hier, um Menschen zu ersetzen, sondern ihnen zu helfen.« Auch »Diella« ist damit eine Verbindung aus KI-gepowertem System und Deepfake-Avatar. Sie verdeutlicht ein bereits langes existierendes Paradoxon zum Thema KI: Einerseits ist das Vertrauen in KI bei

Die virtuelle Ministerin für öffentliche Beschaffung Albaniens.

vielen Menschen geringer als zu anderen Menschen, andererseits aber werden viele (vor allem abstrakte und übermenschliche) Hoffnungen (hier: Integrität) auf Maschinen und Technologie projiziert.

KI-Desinformation

KI-Desinformation kann in verschiedene Formen unterteilt werden, die sich in ihrer technischen Ausprägung und in ihrer Absicht deutlich unterscheiden. Zu den wichtigsten gehören:

1. Fake-News-Webseiten
Hierbei handelt es sich um Internetseiten, die mithilfe von KI erstellte Artikel, Bilder oder audiovisuelle Inhalte veröffentlichen, deren Inhalt ganz oder teilweise frei erfunden ist. Meist werden diese Inhalte in großer Menge produziert und verbreitet.

2. KI-Bilder
KI-generierte Bilder überschwemmen zunehmend soziale Netzwerke, Messenger-Dienste und Nachrichtenportale. Sie werden auf vielfältige Weise eingesetzt: als illustrierende

Begleitung zu Beiträgen in sozialen Medien, als Bebilderung von Artikeln auf Desinformations- oder Clickbait-Webseiten, als visuelle Elemente in Videos oder als Profilbilder falscher Accounts. Häufig zeigen sie Personen (insbesondere Politiker, Prominente, Journalistinnen und Influencer) in Situationen, die nie stattgefunden haben, oder sie stellen Ereignisse dar, die nie geschehen sind. Beispiele sind etwa KI-Bilder, die Taylor Swift angeblich im Wahlkampf für Donald Trump zeigen, oder manipulierte Aufnahmen, die erfundene Kriegszerstörung und Opfer im Gaza-Konflikt darstellen.

3. Deepfakes
Deepfakes sind von KI erzeugte oder manipulierte Video- und Audioinhalte. Der Begriff setzt sich aus »deep learning« und »fake« zusammen. Es gibt verschiedene Arten von Deepfakes, die sich in ihrer Nutzung oder in ihrer Absicht unterscheiden. Dazu gehören etwa Gesichtstausch-Videos, die für pornographische Zwecke, Betrugsversuche oder politische Manipulation verwendet werden. Andere Deepfakes dienen der gezielten Rufschädigung oder der Täuschung in Wahlkämpfen. Diese Inhalte wirken besonders überzeugend, weil sie technisch immer realistischer werden, schwer zu erkennen sind und häufig das Vertrauen der Öffentlichkeit in visuelle Beweise erschüttern. Die ständige Weiterentwicklung der Technologie erschwert auch spezialisierter Software die zuverlässige Erkennung, was Unsicherheit und Misstrauen in der Bevölkerung verstärkt.

4. KI-generierte und automatisierte Beiträge in sozialen Medien
Akteure, die Desinformation verbreiten, nutzen KI-Anwendungen – häufig große Sprachmodelle wie ChatGPT – um massenhaft kurze Beiträge oder Kommentare zu erstellen. Diese werden mit geringen Abwandlungen auf sozialen Plattformen veröffentlicht, etwa um politische Positionen zu unterstützen, Gegner zu diskreditieren oder Themen künstlich in

KI-generiertes Bild (ChatGPT)

Ein junger Mensch steht mit gesenktem Blick vor einer Wand, die vollständig mit bedrohlichen Cybermobbing-Sprüchen und verletzenden Kommentaren bedeckt ist.

den Vordergrund zu rücken. Oft stammen die Beiträge von gefälschten Profilen, die von den gleichen Akteuren gesteuert werden.

5. KI-Übersetzungen
Auch Übersetzungsfunktionen generativer KI werden gezielt für Desinformationskampagnen eingesetzt. Sie ermöglichen es, Beiträge, Artikel, Webseiten, E-Mails oder ganze Drehbücher schnell und in vielen Sprachen zu verbreiten. Untersuchungen von OpenAI zeigten, dass staatliche und private Akteure aus Russland, Iran, China und anderen Ländern ChatGPT nutzten, um während Desinformations- und Cyberangriffen Texte in verschiedene Sprachen zu übersetzen und so ihre Reichweite zu vergrößern.

6. KI-gesteuerte Fake-Profile und automatisierte Bots
Sprachmodelle können vollständige Social-Media-Profile betreiben, die täuschend echt wirken. Solche Profile posten,

kommentieren, liken und teilen Inhalte eigenständig und können dank der Kombination von Text- und Bildgeneratoren sehr überzeugend auftreten. Forschungen zeigen, dass manche dieser Accounts vollständig automatisiert sind, während andere von Menschen geführt werden, die KI-Inhalte kopieren und weiterverbreiten. Diese hybride Struktur erschwert die Unterscheidung zwischen echten Nutzern, menschlich gesteuerten »Trollen« und vollständig KI-gesteuerten Profilen. Diese Form der automatisierten Kommunikation wird im Abschnitt zu »KI-Agenten« näher untersucht.

7. Deepfake Defence und Liar's Dividend

Diese Begriffe bezeichnen eine rhetorische Strategie, bei der politische, wirtschaftliche oder juristische Akteure wahre Informationen als angeblich manipuliert darstellen, um ihre Glaubwürdigkeit zu untergraben. Der bloße Hinweis, ein Bild oder Video könne mithilfe von KI verändert worden sein, reicht dabei oft aus, um Zweifel zu säen und Beweise in Frage zu stellen. Diese Form von Manipulation zeigt, dass Desinformation nicht nur durch KI erzeugt, sondern auch durch den bloßen Verdacht bzw. Verdächtigung auf ihren Einsatz verstärkt werden kann. Alleine die Existenz und die Potentiale von KI genügen also, um Menschen dazu zu bringen, die Echtheit und Glaubwürdigkeit von Informationen grundsätzlich infrage zu stellen. Diese wachsende Unsicherheit wird gezielt für Manipulationen ausgenutzt. Akteure, die Desinformation verbreiten, nutzen das entstandene Misstrauen, um Informationen, Personen und Institutionen anzugreifen, Verwirrung zu stiften und Zweifel zu säen. Die sogenannte Deepfake Defence ist damit keine technische, sondern eine kommunikative Manipulation. Sie wird von Politikerinnen, Juristen, Unternehmern und anderen Akteuren eingesetzt, um belastende Informationen oder Beweise zurückzuweisen, mit dem Hinweis, diese könnten durch generative KI verändert worden sein. Die eigentliche Täuschung geschieht dabei nicht durch Technologie, sondern in der Wahrnehmung: Der bloße Verweis auf die Existenz

KI-generiertes Bild (ChatGPT)

Eine Gegenüberstellung zweier Personen, um einen Face-Swap zu demonstrieren: Links jeweils das Originalporträt der beiden Figuren, rechts die manipulierte Version.

von KI genügt, um Zweifel zu erzeugen und Glaubwürdigkeit zu untergraben, selbst wenn gar keine KI zum Einsatz kam. Beispiele aus der jüngeren Vergangenheit zeigen, wie wirksam diese Strategie sein kann. Während der Wahlen in Mauritius im Jahr 2024 nutzte die Regierung die Deepfake Defence, um die Bedeutung eines abgehörten Telefonskandals herunterzuspielen. Unter dem Vorwand, Deepfakes würden im Umlauf sein, wurden soziale Medien zeitweise blockiert. In Gabun beriefen sich 2019 oppositionelle und militärische Akteure auf denselben Vorwurf, um ihren Putsch gegen Präsident Ali Bongo zu rechtfertigen. Sie erklärten seine Neujahrsansprache für ein Deepfake und behaupteten, er sei tot und daher nicht regierungsfähig. Auch in den Vereinigten Staaten wurde die Deepfake Defence in Gerichtsverfahren angewandt: Anwälte von Donald J. Trump und Elon Musk verwiesen in Prozessen darauf, dass vorgelegte Beweise angeblich KI-generiert sein könnten. Im sogenannten Superwahljahr 2024 griffen rechtspopulistische Parteien und Bewegungen in Europa immer wieder zu dieser Argumentation. In Deutschland behaupteten führende Vertreter der extremen Rechten, Medien hätten die Zahl der Teilnehmenden an Demonstrationen gegen Rechtsextremismus mithilfe KI-manipulierter Bilder übertrieben dargestellt.

Solche Fälle zeigen, dass der Schaden nicht allein durch gefälschte Inhalte entsteht, sondern auch durch falsche Behauptungen über deren Existenz. Je mehr synthetische Inhalte online zirkulieren, desto leichter wird es, wahre Informationen als potenziell künstlich zu diskreditieren. KI stellt die Informationssicherheit somit auf doppelte Weise infrage: Sie kann Inhalte fälschen, und sie kann den Glauben an echte Inhalte erschüttern. Dadurch fördert sie Misstrauen, Unsicherheit und Zweifel an der Verlässlichkeit von Fakten und Wahrheiten.

Was macht KI mit Desinformation?

Dieser Frage wurde in den vergangenen Jahren viel Aufmerksamkeit geschenkt. In der wissenschaftlichen Diskussion stehen sich zwei Positionen gegenüber: Die eine warnt vor einem

»Informationsapokalypse«-Szenario, in dem Deepfakes und KI-generierte online Inhalte das grundsätzliche Vertrauen in Informationen komplett zerstören und online Nutzer noch viel stärker als bislang manipuliert werden können. In der anderen Position wird davon ausgegangen, dass KI das bestehende Desinformationssystem nur wenig verändert. Die Aufmerksamkeit von Zielgruppen sei ohnehin schon voll ausgeschöpft. Deshalb gebe es nur wenig Bedarf an KI-Desinformation.

Während beide Positionen starke Argumente haben, bleiben sie unvollständig: Die erste Position betont vor allem das theoretische Potential (»perfekte Deepfakes zur Massentäuschung«) und theoretische Risiken. Diese Position stand grundsätzlich bereits vor dem großen KI-Hype seit 2022 und betrachtet daher weniger die tatsächliche Realität; die zweite Position vernachlässigt hingegen die Tatsache, dass das Ausmaß KI-basierter Desinformation einerseits kaum genau messbar ist. Hier wird also theoretisch die Notwendigkeit von KI für Desinformation in Zweifel gezogen, während die Praxis klar zeigt, dass sie von den Urhebern und Verbreitern von Desinformation trotzdem eingesetzt wird.

Aufgrund der extrem schnellen Weiterentwicklung von KI-Technologie, ist es schwer, eine definitive und vor allem langfristig gültige Antwort darauf zu geben, wie KI Desinformation verändert. Die meisten Befürchtungen kreisen oft um die Qualität von Fälschungen (insb. Deepfakes) und deren dadurch gesteigerte Überzeugungskraft. Gerade diese Befürchtungen, die mit der neuen Qualität von generativer KI zusammenhängen, haben sich bislang jedoch nicht voll bewahrheitet. Hyperrealistische Fakes haben z. B. 2024 für sich alleine keine Wahlen entscheidend beeinflussen können. Dabei hat sich auch gezeigt, dass hochqualitative und realistische Deepfakes nur kurzfristig, z. B. im Zuge von live Videocalls, überzeugen können, jedoch viel weniger langfristig wirken können.

Im Gegensatz dazu belegen zahlreiche Fallstudien inzwischen, dass generative KI gezielt eingesetzt wird, um Desinformationskampagnen zu vergrößern und zu beschleunigen.

Diese Beispiele zeigen, dass die eigentliche Stärke von KI nicht in der Überzeugungskraft einzelner Inhalte liegt, sondern in ihrer Fähigkeit, Produktion und Verbreitung von Desinformation zu erleichtern. Konkret betrifft dies die digitale Infrastruktur von Desinformation (z. B. Webseiten, bots, fake profile, falsche Profilbilder), automatische KI-generierte Übersetzungen oder einfach generierte Posts, Kommentare, Hashtags etc. Die Quantität, also das Ausmaß und die Anzahl, von Desinformation lässt sich durch KI-Technologien also viel leichter steigern als ihre Qualität. Dauerbeschallung durch KI-gepowerte automatisierte social media Profile ist damit also viel leichter und viel kostengünstiger möglich als spezielle Einzelkampagnen. Quantität und Reichweite manipulativer Botschaften kann so erheblich gesteigert werden. Gleichzeitig senkt KI die Kosten, den Aufwand und auch das Risiko für die Urheber und Verbreiter von Desinformation. Diese können dadurch nicht nur Geld einsparen bzw. mehr Geld verdienen, sondern die Eintrittsbarrieren und das Risiko für sie ist ebenfalls erheblich gesunken.

Eine andere Eigenschaft von KI-Technologie, die sich für die Verbreitung von Desinformation gut eignet und vor allem Masse und Reichweite verändert, ist ihre Fähigkeit zur Datenanalyse und deren Umsetzung in konkrete Inhalte. KI-Systeme können die Funktionsweise von Verteilungs- und Empfehlungssystemen sozialer Netzwerke analysieren und sich gezielt an deren Algorithmen oder auch bestehende Trends und Hashtags anpassen. Sie sind in der Lage, gleichzeitig in verschiedenen digitalen Echokammern aktiv zu sein, passende Hashtags und Schlüsselbegriffe zu erzeugen und Inhalte so zu gestalten, dass sie von Such- und Empfehlungsmechanismen bevorzugt werden. Zudem kann KI visuelle Inhalte und scheinbare Beweise so anpassen, dass sie bestehende Emotionen und Überzeugungen bestimmter Gruppen ansprechen und verstärken. Damit verschiebt sich der Fokus von der Qualität der Desinformation hin zu ihrer Geschwindigkeit, Reichweite und strategischen Wirkung.

KI-Agenten und automatische Desinformation

KI-Agenten und agentische KI-Systeme sind ein nächster Schritt in der Entwicklung und Evolution von KI. Diese Anwendungen können nicht nur Inhalte erzeugen, sondern auch eigenständig Handlungen und Aufgaben ausführen. Damit eröffnet sich auch eine neue Dimension der Informationsmanipulation. Akteure, die gezielt Desinformation verbreiten, setzen agentische KI bereits ein, um die Erstellung manipulativer Inhalte direkt mit deren automatischer Verbreitung zu verknüpfen. Mithilfe solcher Systeme lassen sich gefälschte Social-Media-Profile betreiben, die selbstständig posten, teilen, kommentieren und liken. Diese Form der Automatisierung senkt die Kosten und steigert gleichzeitig Geschwindigkeit und Reichweite manipulativer Kampagnen.

Wie bei anderen Formen von KI-Desinformation besteht auch hier das Problem der Erkennbarkeit. Agentische KI hinterlässt kaum Spuren, und ihr tatsächliches Ausmaß ist schwer zu bestimmen. Hinweise auf ihren Einsatz ergeben sich meist nur durch identisches Verhalten mehrerer Accounts, die gleichzeitig und inhaltlich abgestimmt agieren. Studien zeigen, dass solche Systeme bereits aktiv genutzt werden. Das Unternehmen Anthropic, bekannt durch seinen Chatbot »Claude«, veröffentlichte im April 2025 einen Bericht, in dem es einen »Disinformation-as-a-Service«-Akteur (eine PR-Agentur in den Philippinen) identifizierte. Dieser betrieb über hundert verschiedene Social-Media-Personas auf X und Facebook und schuf so ein Netzwerk politisch ausgerichteter Profile, die mit zehntausenden echten Nutzern interagierten. Laut Anthropic nutzte die Gruppe Claude, um taktische Entscheidungen zu treffen, etwa, ob ein Account einen Beitrag liken, teilen, kommentieren oder ignorieren sollte, und um Prompts für Bildgeneratoren zu erstellen und deren Ergebnisse zu bewerten. Ziel war es, politische Narrative zu verbreiten, die bestimmte europäische, iranische und kenianische Interessen unterstützten oder schwächten.

In einem anderen Beispiel dokumentierte die Organisation »AI Forensics« den Einsatz agentischer KI durch private Nutzer.

KI-generiertes Bild (ChatGPT)

Eine Person untersucht mit einer Lupe, die als KI-Detektor fungiert, ein Bild, um die Verwendung von Künstlicher Intelligenz zu erkennen.

Diese verwenden halbautomatische Systeme, um massenhaft minderwertige oder falsche KI-Inhalte zu verbreiten, darunter »Deepfake-Prominente«, erfundene historische Darstellungen oder Verschwörungstheorien, die vor allem auf Plattformen wie TikTok und Instagram für Klickzahlen und Einnahmen sorgen.

Bislang wird agentische KI vor allem für zwei Zwecke eingesetzt: zum automatisierten Betrieb falscher Social-Media-Accounts und zur automatisierten Verbreitung von KI-Falschinhalten über diese Accounts. Damit bringt sie eine neue Dimension von Geschwindigkeit, Quantität und Effizienz in die Welt der Desinformation. Besonders wirksam ist sie durch

ihre Fähigkeit, Text, Bild und Ton zu kombinieren, was sie optimal an die multimodalen Anforderungen sozialer Plattformen anpasst. Diese Eigenschaften machen agentische KI sowohl für kommerzielle Akteure, die Desinformation als Dienstleistung betreiben, als auch für staatlich geförderte Einflussoperationen attraktiv, die Informationsräume gezielt überfluten und destabilisieren wollen.

Beispiele

Beispiel 1: Copycop / Storm1516
CopyCop, auch bekannt unter dem Codenamen *Storm-1516*, bezeichnet ein umfangreiches russisches Einflussnetzwerk, das von John Mark Dougan, einem ehemaligen US-Polizisten, aufgebaut und betrieben wird. Dougan flüchtete 2016 nach Russland und hat sich dort zu einem zentralen Akteur in der Kreml-Propaganda entwickelt. Laut Erkenntnissen von US-Behörden wird Dougan vom russischen Militärgeheimdienst GRU angeleitet, finanziert und unterstütz. Das CopyCop-Netzwerk besteht aus Hunderten gefälschten Nachrichtenwebsites und Social-Media-Accounts, die den Anschein lokaler Medien oder Graswurzel-Initiativen erwecken. Seit 2024 hat das Netzwerk sein Operationsgebiet rasant ausgedehnt: Waren 2024 nur ca. 80 englischsprachige und auf die US-Wahlen fixierte Fake-Seiten bekannt, kamen um die deutsche Bundestagswahl 2025 über 100 weitere deutschsprachige Seiten hinzu bis die Gesamtzahl Ende 2025 auf über 300 Fake News Websites in verschiedenen Sprachen, die sich speziell auf die USA, Kanada, Westeuropa und sogar Afrika richten. Diese Seiten geben vor, Nachrichtenportale, politische Parteien oder Faktenchecker zu sein, um pro-russische Desinformation glaubwürdiger zu verbreiten. Beispielsweise flog auf, dass CopyCop-Betreiber im Juni 2024 die Domain einer französischen Regierungspartei imitierten, um Falschinfos im Kontext der Europawahl zu verbreiten.

Die strategischen Ziele von CopyCop decken sich weitgehend mit der offiziellen Linie des Kremls: Unterminierung der

westlichen Unterstützung für die Ukraine, Destabilisierung westlicher Demokratien durch das Schüren und Anheizen gesellschaftlicher Konflikte sowie Förderung russischer Interessen in Einflusszonen. Dazu verbreitet das Netzwerk laufend propagandistische Artikel und Videos mit stark anti-ukrainischen, anti-westlichen Narrativen auf seinen Fake-Seiten. Beliebte Themen und Botschaften sind dabei die angebliche Ineffektivität westlicher Sanktionen, Korruptionsvorwürfe gegen ukrainische Politiker oder Verschwörungsthesen über westliche Politiker oder Migration. So lancierte Storm-1516 etwa die falsche Story, US-Vizepräsidentin Kamala Harris habe 2011 einen Verkehrsunfall verursacht und Fahrerflucht begangen oder streute erfundene Missbrauchsvorwürfe gegen den US-Politiker Tim Walz. In Deutschland wurden prominente Regierungsmitglieder wie Friedrich Merz, Robert Habeck oder Annalena Baerbock mit frei erfundenen Skandalen – von Korruption bis hin zu bizarren Verschwörungen – diffamiert. Diese Narrativen werden *multimedial* aufbereitet: CopyCop produziert Kurzvideos mit angeblichen »Whistleblowern«, fingierte Interviews sowie lange Dossiers, die angebliche Beweise präsentieren sollen. Typisch ist auch das Erstellen kompletter Fake-Nachrichtenseiten, deren Design und Schreibstil realen Medien nachempfunden sind. Dougan und seine Mitstreiter koordinieren zudem ein Netzwerk von pro-russischen Influencern und Trollen in sozialen Medien, die die Inhalte der Fake-Websites aktiv weiterverbreiten. Laut Untersuchungen erzielen CopyCop-Beiträge mit hohen Reichweiten. Bisweilen schaffen sie es auch in öffentliche Debatten einzudringen.

Das CopyCop-Netzwerk setzt in erheblichem Umfang KI ein, um breiter, effizienter und glaubwürdiger zu arbeiten. Ein zentrales Element ist die Nutzung von KI-Chatbots (LLMs) zur automatisierten Texterstellung. Bereits 2024 wurde bekannt, dass CopyCop-Betreiber Artikel etablierter Nachrichtenportale plagiieren und sie mithilfe von KI umschreiben, um einen pro-russischen Spin hinzuzufügen. Diese Praxis ermöglicht es, in kurzer Zeit zahlreiche vermeintlich journalistische

Beiträge zu generieren, die auf den ersten Blick seriös wirken, tatsächlich aber subtil Desinformation verbreiten. Nach eigenen Angaben sprach Dougan davon, rund 90.000 Artikel monatlich per KI generieren und veröffentlichen zu lassen. Allerdings stießen er und sein Team dabei auch auf Hindernisse: Eigenen öffentlichen Aussagen Dougans zufolge eigneten sich diese Modelle nur bedingt für russische Desinformation, da sie auf westlichen Trainingsdaten basieren. Dies führte dazu, dass die generierten Texte aus Sicht der Propagandisten nicht russslandfreundlich oder propagandistisch genug ausfielen, unerwünschte Neutralität zeigten oder pro-Russische Inhalte nicht gut genug für ein westliches Publikum aufbereiten konnten. Die Reaktion der CopyCop-Architekten war es, vom anfangs genutzten ChatGPT auf speziell trainierte, ungefilterte und lokal gehostete Open Source Modelle umzusteigen. Seit 2025 nutzt Dougan deshalb in seiner Wohnung selbst gehostete Modelle, die auf Meta's LLaMA 3-Architektur basieren. Konkret identifizierten die Forscher zwei angepasste Varianten von Llama-3-8B, die ohne Inhaltsfilter frei propagandistische Texte erzeugen können. Russische Stellen sollen sogar aktiv die Infrastruktur für diese KI bereitgestellt haben, laut Berichten finanzierte die GRU die nötigen Server, während das CGE bei der Entwicklung von Deepfake-Inhalten mithalf. Dougan selbst beklagte in einem Moskauer Forum im Januar 2025, dass es derzeit keine wirklich guten KI-Modelle gibt, um russische Nachrichten hochzuskalieren, und forderte, man müsse eigene, russische Modelle ohne westliche Voreingenommenheit speziell mit russischen Medienarchiven trainieren.

Diese Aussagen unterstreichen, wie wichtig der KI-Einsatz für CopyCop geworden ist. Durch den KI-Einsatz konnte CopyCop sein Propagandanetz exponentiell ausbauen. Während ein traditionelles Trollfabrik-Modell (wie einst die Internet Research Agency) hunderte Mitarbeiter benötigte, um Inhalte per Hand zu erstellen, erledigt nun die KI einen Großteil der Texterstellung und Übersetzung automatisch. Dies erklärt, warum innerhalb weniger Monate über 200 neue Fake-Seiten mit

Inhalten in Türkisch, Ukrainisch, Swahili und weiteren Sprachen online gehen konnten. KI-gestützte Automatisierung hält die Fake-Portale zudem aktiv: Viele der CopyCop-Seiten publizieren ständig generische Artikel mit pro-russischer Stoßrichtung, die komplett von Sprachmodellen generiert sind. Diese dienen als Füllmaterial, um die Seiten lebendig wirken zu lassen, bis gezielt auf einzelne Ereignisse zugeschnittene Desinformationen eingestreut werden. Darüber hinaus kommt KI-basierte Bild- und Videomanipulation zum Einsatz. Das Netzwerk erstellt überzeugende Deepfakes, von gefälschten Video-Interviews mit angeblichen Whistleblowern bis hin zu manipulierten Audioaufnahmen, in denen die Stimmen westlicher Prominenter oder Politiker imitiert werden. Insgesamt macht erst der umfangreiche KI-Einsatz das CopyCop/Storm-1516-Netzwerk zu einer neuen Generation der

KI-generiertes Bild (ChatGPT)

Im Inneren der Matrjoschka-Puppe leuchtet ein grüner Binärcode, außen sind Zeitungstexte und politische Symbole zu sehen – Sinnbild für digitale Desinformation und Propaganda.

Desinformationskampagne. KI dient hier als Hebel, um mit begrenztem Personalaufwand maximale Wirkung zu erzielen: automatische Texterstellung in vielen Sprachen, schnelle Skalierung auf neue Themenfelder und audiovisuelle Fälschungen.

Beispiel 2: Operation Overload

Operation Overload ist eine mindestens seit 2023 aktive, prorussische Desinformationskampagne, die durch ihr Volumen und ihre Täuschungsmethoden auffällt. Ihr Hauptziel besteht darin, westliche Medien, Faktenchecker, Experten und Forscher mit gefälschten Inhalten zu überfluten und ihre Arbeit dadurch zu behindern. Dies wiederum soll andere breitflächige russische Desinformationsoperationen begünstigen. Charakteristisch für *Overload* ist das systematische »Überlasten« Faktenchecker, Medien, Journalisten und Forschern: Hunderte Medienhäuser in Europa, den USA und anderen NATO-Staaten erhalten so regelmäßig massenhaft E-Mails und Social-Media-Nachrichten mit Aufforderungen, erfundene Behauptungen oder manipulierte Bilder zu verifizieren. Diese Inhalte zielen vor allem darauf ab, die Ukraine und pro-westliche Regierungen zu diskreditieren, westliche Hilfsbereitschaft zu unterminieren und gesellschaftliche Spannungen in Demokratien zu schüren. *Overload* arbeitet also mit raffinierten Imitations- und Täuschungstaktiken, um westliche Abwehrversuche gegen Desinformation unwirksam zu machen. In koordinierten Schüben werden auf verschiedenen online Plattformen wie Telegram, X (Twitter) und Bluesky falsche Nachrichtenbeiträge, gefälschte Dokumentationen und vermeintliche Leaks veröffentlicht, die seriöse Quellen imitieren. Die Akteure erstellen Fake-Profile, die vorgeben, Journalisten, Wissenschaftler oder Behörden zu sein, und statten diese mit gestohlenen Logos, manipulierten Bildern und erfundenen Zitaten aus.

Im Sommer 2024 kursierten z. B. auf Telegram Pseudo-TV-Dokumentationen mit dem Titel »Olympics Has Fallen«, in denen mit KI generierte Videos die Sicherheit der Pariser Olympischen Spiele infrage stellten. Parallel wurden gefälschte

Titelseiten bekannter Zeitungen und Magazine erstellt, die erfundene Artikel mit echten Markenlogos präsentierten. Diese mehrschichtige Vorgehensweise (das sogenannte *Content-Amalgamierung*-Prinzip) soll den Eindruck erwecken, als gebe es eine breite virale Verbreitung der Falschinformationen auf unterschiedlichen Kanälen. Anschließend verschickten die Drahtzieher gezielte E-Mails an Faktenchecker und Redaktionen, in denen sie auf diese manipulierten Inhalte verwiesen (oft per Link oder QR-Code) und um Überprüfung baten. Ziel war es, die begrenzten Ressourcen der Faktenprüfer zu binden und sie zugleich dazu zu verleiten, über die Fake-Inhalte zu berichten oder Widerlegungen zu veröffentlichen. Ironischerweise tragen solche Debunking-Artikel oft ungewollt dazu bei, die Desinformation weiter zu verbreiten, weil sie die erfundenen Geschichten einem größeren Publikum bekannt machen.

In der Operation Overload spielt der Einsatz von KI mittlerweile eine zentrale Rolle und verleiht der Kampagne eine neue Qualität. Die Drahtzieher verwenden moderne generative KI-Tools, um echt wirkende Medieninhalte in großer Zahl automatisch zu erzeugen. Ein prägnantes Beispiel ist die Verwendung von KI-generierten Stimmen, um Audioaufnahmen und Videos zu erstellen, die klingen, als stammten sie von realen Journalisten, Professoren oder Polizeisprechern. Laut Untersuchungen nutzen Overload-Akteure Video-Clips, die echten Nachrichtenbeiträgen nachempfunden sind, und versehen sie mit synthetischen Sprecherstimmen, die bekannten Experten gehören sollen. Auf diese Weise wurden im ersten Quartal 2025 über 80 renommierte Institutionen und Medien nachgeahmt, indem man entweder ihre Logos unautorisiert einblendete oder die Stimmen ihrer Vertreter per KI-Klon nachahmte. Gleichzeitig wurden Bilder und Überschriften gefälscht, teils durch Bildbearbeitung, teils durch KI-Bildgeneratoren, um Schlagzeilen vertrauenswürdiger Nachrichtenportale vorzutäuschen. KI-Technologie ermöglicht es Overload so, mit relativ geringem personellen Aufwand mehr und vielfältigere falsche Inhalte zu produzieren und deren Qualität zu erhöhen. Think

Tanks und Experten beobachteten seit Ende 2024 einen »alarmierenden Anstieg sowohl in der Menge als auch der Komplexität koordiniert verbreiteter Falschinhalte«, die häufig KI-generiert oder manipuliert waren. Ohne generative Sprach- und Bildmodelle wäre es kaum machbar, innerhalb weniger Monate Hunderte einzelne Fake-Beiträge in bis zu zehn Sprachen zu erstellen und über diverse Kanäle auszuspielen. KI bietet hier also einen Multiplikator-Effekt: Sie erhöht die Geschwindigkeit, Skalierbarkeit und Glaubwürdigkeit der Desinformation. So können per KI erstellte Stimmen und Gesichter deutlich überzeugender wirken als reine Text-Täuschungen, was die Chancen erhöht, dass unwissende Zuschauer die Fälschungen für echt halten. Ferner senken frei verfügbare KI-Tools die Kosten und Eintrittsbarrieren. Genau das macht Overload aus: Die Kampagne setzt allgemein zugängliche KI-Werkzeuge strategisch ein, um eine Flut von Desinformationen zu erzeugen, die Faktenprüfer und Medien wortwörtlich überlasten soll. Faktenchecker und Medien berichteten von einer Flut von bis 800 offenbar KI-generierten automatischen Emails pro Tag, die auf gefälschte Inhalte aufmerksam machen wollen. Dadurch unterscheidet sich Overload von früheren Desinformationsoperationen: Erstmals kommt KI in diesem Ausmaß zum Tragen, um Informationswächter gezielt in die Irre zu führen und deren Arbeit erheblich zu erschweren.

KI-generierte Inhalte: Was sagt das Gesetz?

In Europa, aber auch weltweit wurden in den vergangenen 2 Jahren mehrere Gesetze und Richtlinien verabschiedet, die die Gefahren von KI und KI-Inhalten mindern sollen. Für die Europäische Union ist hier vor allem der EU AI Act (»KI-Gesetz der EU«) zu nennen, das weltweit erste umfassende Regelwerk zur Regulierung Künstlicher Intelligenz. Er legt fest, wie KI-Systeme nach ihrem Risikoniveau eingestuft und kontrolliert werden, und klassifiziert Deepfakes in sensiblen Bereichen wie Medien, Werbung und politischen Kampagnen als Hochrisikoanwendungen. Für sie gelten besonders strenge Aufsichts- und Transparenzpflichten, um Manipulation und Desinformation zu verhindern. Der AI Act untersagt ausdrücklich den Einsatz von KI-Anwendungen, die Menschen täuschen oder manipulieren, etwa Deepfakes, die falsche Informationen verbreiten oder den Ruf von Personen schädigen. Zudem schreibt er vor, dass alle von KI erzeugten oder veränderten Inhalte klar als solche gekennzeichnet werden müssen. Verstöße gegen diese Vorgaben können mit erheblichen Geldstrafen geahndet werden (bis zu 35 Millionen Euro oder sieben Prozent des weltweiten Jahresumsatzes eines Unternehmens).

In Bezug auf Desinformation und damit politische Manipulation mit KI, verweist der AI Act auf ein anderes, bereits bestehendes europäisches Gesetz, den Digital Services Act (DSA). Dieser bildet den Kern der EU-Strategie gegen Desinformation und schafft den regulatorischen Rahmen für Online-Plattformen. Der AI Act ergänzt den DSA, der insbesondere sehr große Online-Plattformen verpflichtet, Desinformation und illegale Inhalte aktiv zu entfernen. Artikel 40 des DSA gewährt geprüften Forschern Zugang zu öffentlichen und nichtöffentlichen Daten dieser Plattformen, um die systemischen

Risiken von Desinformation besser zu untersuchen und eine wirksamere Kontrolle zu ermöglichen. Bei Verstößen drohen Plattformen Bußgelder von bis zu sechs Prozent ihres Jahresumsatzes. Diese Regelungen gelten sowohl für Desinformation, die mithilfe von KI erzeugt wird, als auch für nicht KI-basierte Formen.

Im Rechtsraum der EU setzt der Gesetzgeber also in Bezug auf KI-generierte Inhalt also einerseits auf eine generelle Kennzeichnungspflicht von KI-Inhalten und transparente Regeln für deren Verbreitung auf online Plattformen und andererseits auf Verbote bestimmter Inhalte (Deepfakes und Deep porn). Die Umsetzung dieser Regeln legt der Gesetzgeber ferner – innerhalb bestimmter Fristen – den Plattformen selbst auf. Diese müssen also in den nächsten Jahren automatische Erkennungs- und Kennzeichnungsmechanismen etablieren und ihre Inhaltsmoderation entsprechend anpassen. Genau darum tobt jedoch eine anhaltende, heftige politische und rechtliche Auseinandersetzung.

Doch auch außerhalb des KI-Gesetzes der EU gibt es viele Gesetze, Richtlinien und Vereinbarung, die verschiedene Formen von KI-Manipulation behandeln. Viele Länder, darunter die USA, verschiedene US-Bundesstaaten, das Vereinigte Königreich und viele weitere, haben zum Bespiel spezielle Gesetze erlassen, die einerseits die Erstellung und andererseits die Verbreitung von Deepfake-Pornographie unter Strafe stellen. Andere Gesetze, z. B. im US-Bundesstaat Kalifornien, regeln speziell Kennzeichnungspflichten und auch Verbote KI-generierter Wahlwerbung. Dänemark hingegen erließ ein spezielles Gesetz, dass die Persönlichkeitsrechte jedes Individuum vor missbräuchlichen Anwendungen in Deepfakes schützen soll.

Im internationalen Rahme sorgte die G7»Hiroshima Declaration on AI« aus dem Jahr 2023 als erste für eine Reihe von Bestimmungen und Empfehlungen, die sich gezielt mit den Risiken von Desinformation und Deepfakes befassten. Sie betont die Notwendigkeit eines umfassenden politischen Rahmens,

KI-generiertes Bild (ChatGPT)

Die Straße führt geradeaus in die Ferne, während links und rechts die aus Paragraphen geformten Schutzleitplanken die Richtlinien vorgeben.

der sichere und vertrauenswürdige KI-Systeme fördert und gleichzeitig die mit generativer KI verbundenen Risiken im Bereich von Falsch- und Desinformation mindert. Die Erklärung hebt Transparenz als zentrales Prinzip hervor: Organisationen, die fortgeschrittene KI-Systeme entwickeln, sollen ihre Inhalte klar als KI-generiert kennzeichnen, um Nutzerinnen und Nutzern eine bessere Bewertung der Informationsquelle und -authentizität zu ermöglichen. Darüber hinaus fordern die G7-Staaten eine verstärkte internationale Zusammenarbeit bei der Entwicklung von Verfahren zur Erkennung und Bekämpfung von Deepfakes sowie die Förderung gemeinsamer Forschungsinitiativen zur Bekämpfung von Desinformation. Die Erklärung unterstreicht außerdem die Bedeutung eines kooperativen Ansatzes, der Regierungen, Wissenschaft und Zivilgesellschaft gleichermaßen einbindet.

Auch die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) hat sich mit den Themen KI-Desinformation und Deepfakes befasst, insbesondere in ihren

AI Principles, der OECD Recommendation of the Council on Artificial Intelligence und dem OECD Framework for the Classification of AI Systems. Die OECD AI Principles betonen die Bedeutung von Transparenz und Verantwortlichkeit in der Entwicklung und Anwendung von KI. Sie sprechen sich für eine klare Kennzeichnung KI-generierter Inhalte aus, um Fehlinformationen vorzubeugen, und betonen die Notwendigkeit robuster und sicherer Systeme, die Missbrauch – etwa durch die Erstellung manipulativer Deepfakes – verhindern sollen. Die OECD-Empfehlung des Rates zur Künstlichen Intelligenz, die 2024 aktualisiert wurde, behandelt das Thema Informationsintegrität bei generativer KI. Sie fordert verantwortungsbewusstes Handeln von Unternehmen und legt besonderen Wert auf Transparenz hinsichtlich der Fähigkeiten, Grenzen und Datenquellen von KI-Systemen, um Risiken durch Desinformation einzudämmen. Das OECD »Framework for the Classification of AI Systems« schließlich bietet politischen Entscheidungsträgern eine strukturierte Methode, um KI-Systeme nach ihrem möglichen Einfluss und Risiko – einschließlich des Risikos der Desinformation – einzuordnen und zu bewerten.

KI-Inhalte: Was sagen die Social Media Plattformen?

Soziale Medien sind die Hauptverbreitungsorte von KI-generierten Inhalten und werden deshalb nicht nur in Gesetzen wie dem KI-Gesetz der EU speziell behandelt. Viele online Plattformen und Messenger-Dienste haben sich – auch als Reaktion auf die Gesetzesregelungen – eigene Regeln (»Community Guidelines« und Nutzungsbedingungen) für KI-Inhalte gegeben oder entsprechende Verweise in bestehende Regeln aufgenommen.

So enthalten *Metas* »Community Guidelines« spezielle Bestimmungen zu KI-generierten Inhalten: Nutzer müssen Inhalte kennzeichnen, die digital erstellt oder verändert wurden und andere in die Irre führen könnten. Seit Mai 2024 begannen Plattformen von Meta mit der automatischen Erkennung von KI-Inhalten, um sie mit dem Hinweis »Made with AI« zu versehen. Darüber hinaus arbeitet Meta mit anderen Unternehmen an einheitlichen technischen Standards zur Identifizierung von KI-Inhalten. In der Praxis bleibt jedoch unklar, wie zuverlässig diese Erkennung funktioniert und ob die Regeln weltweit einheitlich umgesetzt werden. *WhatsApp*, obwohl ebenfalls zu Meta gehörend, erwähnt in seinen Richtlinien weder KI-generierte Inhalte noch Deepfakes oder Desinformation ausdrücklich, sondern verweist nur allgemein auf die unzulässige betrügerische Nutzung des Dienstes. Auch der Messenger-Dienst *Telegram* erwähnt KI- oder Deepfake-Inhalte nicht direkt, sondern verlässt sich darauf, dass Nutzer illegale Inhalte melden.

YouTube verlangt von Inhaltserstellern, den Einsatz von KI-generiertem oder verändertem Material offenzulegen. Für

Videos, die Ereignisse oder Aussagen darstellen, die in Wirklichkeit nicht stattgefunden haben, ist eine Kennzeichnung in der Videobeschreibung verpflichtend. Nutzer können außerdem die Löschung von KI-generierten Inhalten beantragen, die ihr Abbild ohne Zustimmung zeigen. *X* (ehemals *Twitter*) hat seine Richtlinien angepasst und verbietet die Verbreitung von »unauthentischen Medien«, also manipulierten oder aus dem Kontext gerissenen Inhalten, die zu Verwirrung, Sicherheitsrisiken oder ernsthaften Schäden führen können. Allerdings nimmt X in Fällen, in denen eine eindeutige Beurteilung nicht möglich ist, häufig keine Maßnahmen vor. Seit der Integration des KI-Chatbots »Grok« steht die Plattform zudem verstärkt in der Kritik, weil Grok oft selbst falsche Informationen wiederholt sowie manipulative und beleidigende KI-Inhalte und Deepfakes verbreitet hat.

Auch *TikTok* verpflichtet seine Nutzer zur Kennzeichnung sogenannter »synthetischer Medien«. Deepfakes oder veränderte Videos müssen mit entsprechenden Labels, Stickern oder Beschriftungen wie »synthetic«, »AI-generated« oder »not real« versehen werden. Gleichzeitig gilt TikTok als bevorzugte Plattform für Akteure, die generative KI nutzen, um automatisiert massenhaft manipulative Inhalte, verschwörungsideologische Kanäle oder sogenannten »KI-Slop« zu produzieren und damit Geld zu verdienen, oftmals ohne nennenswerte Eingriffe durch die Plattform selbst.

Auf *Reddit* gelten dezentral organisierte Regeln für einzelne »Communities« zum Umgang mit KI-Inhalten. Viele Subreddits haben eigene Richtlinien entwickelt, die von den jeweiligen Moderationsteams festgelegt werden. Eine aktuelle Analyse zeigt, dass zwar noch immer relativ wenige Subreddits explizite Vorgaben zu KI enthalten, ihre Zahl sich im vergangenen Jahr jedoch nahezu verdoppelt hat, besonders in größeren Communities sowie in Foren, die sich mit Kunst oder Prominenten beschäftigen. Die dortigen Regelungen spiegeln häufig Sorgen um Qualität und Authentizität wider, weshalb manche Subreddits KI-generierte Inhalte vollständig verbieten.

In den Community Guidelines von *Discord* finden sich keine gesonderten Regeln zu KI-generierten Inhalten, wohl aber Empfehlungen zum Einsatz von KI im Bereich der Moderation. Discord ermutigt seine Nutzergemeinschaften, KI-gestützte Tools einzusetzen, um unangemessene Inhalte wie Spam, Hassrede oder andere problematische Beiträge zu erkennen und zu verwalten. Desinformation wird dabei allerdings nicht ausdrücklich als eigene Kategorie von unerwünschtem Inhalt genannt. Auch Discord folgt einem dezentralen Prinzip, bei dem einzelne Server ihre eigenen Richtlinien zum Umgang mit KI festlegen können.

Diese Beispiele zeigen, dass die meisten Plattformen KI-Inhalte nicht grundsätzlich verbieten, sondern versuchen, sie in bestehende Regelwerke zu Desinformation und Missbrauch einzuordnen. Das generelle Ziel, im Einklang mit den Anforderungen des EU AI Act, ist die Erkennung und Kennzeichnung von KI-generierten Inhalten. Dabei liegt die Verantwortung für Größe, Platzierung und Gestaltung der Labels meist bei den Plattformen oder den Nutzern. Ab 2026 wird EU-Recht die Erkennung und Kennzeichnung von Deepfakes und KI-Inhalten verpflichtend machen. Bislang verfügen die meisten Anbieter jedoch über keine automatisierten Systeme zur Vorab-Erkennung und setzen weiterhin überwiegend auf nachträgliche Moderation nach Nutzerbeschwerden. Tests haben zudem gezeigt, dass Wasserzeichen, Labels und digitale Herkunftsnachweise leicht entfernt werden können und dass viele Tools zur Deepfake-Erstellung bereits über eigene Funktionen zur Entfernung solcher Markierungen verfügen. Verstöße gegen Kennzeichnungspflichten bleiben in der Regel folgenlos. Trotz der zunehmenden Integration von KI in ihre eigenen Systeme verfolgen die Plattformen daher bislang keinen systematischen oder strategischen Ansatz, um den besonderen Risiken und manipulativen Potenzialen von KI-Desinformation wirksam zu begegnen.

Beispiel: Politische KI-Inhalte auf Social Media und das globale Superwahljahr 2024

2024 war nicht nur ein Jahr der KI, sondern auch ein globales Superwahljahr mit mehr als 50 Wahlen weltweit. Gerade die Befürchtungen vor KI-Manipulationen (insb. Desinformation und Deepfakes) von Wahlen waren deshalb groß. Während der Münchner Sicherheitskonferenz im Februar 2024 unterzeichneten 27 führende Technologieunternehmen, darunter Microsoft, Google, Meta, Amazon, OpenAI und TikTok, den »Munich Tech Accord« (offiziell *Tech Accord to Combat Deceptive Use of AI in 2024 Elections*). Er verpflichtete die Unterzeichner zur Zusammenarbeit bei der Erkennung und Bekämpfung schädlicher KI-Inhalte, insbesondere im Zusammenhang mit Wahlen. Die beteiligten Unternehmen verpflichten sich, vorbeugende Maßnahmen zu entwickeln, etwa durch Forschung und den Einsatz technischer Systeme, die die Verbreitung manipulativer KI-Inhalte einschränken sollen. Sie einigten sich außerdem darauf, generierte Inhalte mit technischen Signalen wie Wasserzeichen zu versehen, um deren Herkunft nachvollziehbar zu machen. Darüber hinaus beinhaltete der Accord gemeinsame Verpflichtungen zur Erkennung und Reaktion auf täuschende Inhalte, insbesondere wenn diese finanziert oder gezielt verbreitet werden. Ein weiterer Schwerpunkt lag auf Aufklärungsarbeit: Bürgerinnen und Bürger sollen durch Bildungsinitiativen und Medienkompetenzprogramme besser auf die Risiken von KI-Manipulation vorbereitet werden.

Viele Plattformen, darunter Meta und Google, haben nicht nur den Munich Tech Accord unterzeichnet, sondern im Jahr 2024 zusätzlich eigene Richtlinien und Nutzungsbedingungen für politische Inhalte veröffentlicht. Diese Regelungen zielten auf politische Werbung und sollten mehr Transparenz schaffen, indem sie eine Kennzeichnung politischer Beiträge und Anzeigen vorschrieben. Unternehmen wie Microsoft und Google haben ihre Informations- und Aufklärungsangebote über KI, Desinformation und Deepfakes erweitert und bieten seit 2024 sogenannte »Prebunking«-Materialien sowie kostenlose

Tools zur Überprüfung von KI-generierten Inhalten an. In der Praxis zeigte sich jedoch, dass diese Maßnahmen weltweit ungleich umgesetzt wurden. Der Schwerpunkt vieler Initiativen lag weniger auf den zahlreichen internationalen Wahlen des Jahres 2024 als vielmehr auf der US-Präsidentschaftswahl, die den größten Teil der Aufmerksamkeit und Ressourcen auf sich zog.

Da 2024 auch die EU-Parlamentswahlen stattfanden, verabschiedete das EU-Parlament zuvor den "Code of Conduct for the 2024 European Parliament Elections". Dieses Abkommen legt einen Rahmen für ethisches Wahlkampfverhalten fest und reagiert insbesondere auf die Herausforderungen, durch KI-generierte Desinformation und manipulative Inhalte. Die Unterzeichner verpflichteten sich, in ihrer politischen Kommunikation den Grundsätzen von Wahrhaftigkeit und Genauigkeit zu folgen und aktiv gegen Falsch- und Desinformation vorzugehen. Der Kodex betonte die ethische Nutzung technologischer Mittel, einschließlich Künstlicher Intelligenz, und fordert die Parteien dazu auf, keine irreführenden Inhalte zu produzieren oder zu verbreiten, die das öffentliche Meinungsbild durch Deepfakes oder andere Manipulationstechniken verzerren könnten. Darüber hinaus stärkt der Kodex die Transparenz, indem er vorschrieb, dass alle Parteien KI-generierte Inhalte in ihren Kampagnen klar kennzeichnen müssen. Dabei handelte es sich um eine freiwillige, rechtlich nicht-bindende Selbstverpflichtung. In der Praxis zeigte dieser Verhaltenskodex sehr wenig Wirkung: Traditionelle und gemäßigte Parteien und Fraktionen benutzten fast gar keine KI und KI-Inhalte in diesem Wahlkampf, wohingegen die rechtsextremen Fraktionen in Frankreich, Italien und Deutschland KI-Bilder, KI-generierte Lieder und KI-Wahlplakate ohne entsprechende Kennzeichnung verwendeten.

KI-generiertes Bild (ChatGPT)

KI erzeugte Bilder kann man an kleinen Fehlern erkennen: Formell gekleidete Menschen durchqueren eine Wüste. Es fehlen jedoch die Fußspuren im Sand.

KI-Erkennung und ihre Tücken

Einer der Gründe für die Angst vor Manipulation durch KI-generierte Inhalte ist ihre neue Qualität. KI erzeugt, kurz gesagt, Inhalte, die mindestens genauso gut sind und aussehen wie die Menschen erzeugten Texte, Bilder oder Videos. Und da die Qualität der dahinterstehenden Programme besser wird, verschwinden Fehler und Erkennungsmerkmale immer schneller, was Menschen verunsichert. KI-Inhalte erkennen, ist deshalb eine »Antwort« auf die Angst vor KI-Manipulationen.

Doch die Erkennung (»detection«) von KI-Inhalten stellt eine Vielzahl von Herausforderungen dar und, so zeichnet sich bereits ab, wird in der Zukunft nur wenig zum Schutz vor Ki-Manipulation beitragen können. Dafür gibt es zahlreiche Gründe:

Einerseits zeigen Studien, dass Menschen deutlich schlechter darin sind, KI-generierte Inhalte zu erkennen als spezialisierte Erkennungsprogramme. Doch auch das ist kein verlässlicher Schutz, denn die Erkennungssoftware selbst arbeitet nie mit hundertprozentiger Treffsicherheit. Ein Grund hierfür ist das sogenannte »GAN-Problem«, also das Wettrennen zwischen Inhalt-generierenden und erkennenden Netzwerken, dass dafür sorgt, dass jedes verbesserte Erkennungsmodell von neuen, besser getarnten KI-Systemen bald überholt wird. Um aktuell zu bleiben, müssen Erkennungsprogramme darüber hinaus ständig mit immer neuen Trainingsdaten gefüttert werden, etwa mit den neuesten Deepfakes oder synthetischen Bildern, um mit der KI-Entwicklung mitzuhalten und nicht zu veralten. Und auch dennoch liefert Erkennungssoftware nicht immer eindeutige Ergebnisse: Statt klarer Aussagen berechnen sie Wahrscheinlichkeiten und präsentieren oft ungenaue, uneindeutige oder auch unbrauchbare Einschätzungen.

Zudem existiert inzwischen eine Vielzahl unterschiedlicher Programme zur KI-Erkennung, deren technische Grundlagen, Trainingsdaten und Bewertungsverfahren für normale Nutzer kaum nachvollziehbar sind. Wer ein Tool nutzt, weiß selten, worauf es eigentlich seine Entscheidung stützt oder wie zuverlässig es wirklich ist. Dazu kommt auch, dass es nicht ein KI-Erkennungstool für alle Inhaltsformen gibt, Nutzer brauchen also KI-Erkennungen für Texte, Bilder, Videos und Audios.

Die Erwartungen an solche Programme sind gleichzeitig enorm hoch: Nutzer verlangen einfache, schnelle und eindeutige Antworten, was technisch nicht zu erfüllen ist. Ebenso ist die Bereitschaft, Geld für KI-Erkennungsprogramme zu investieren und zu bezahlen, immer noch gering. Und hinzu kommt ein weiteres Problem: Solange Erkennungsprogramme nicht direkt in alltägliche Anwendungen wie Browser, E-Mail-Clients, soziale Plattformen oder Videokonferenz-Software integriert sind, bleibt ihr Nutzen begrenzt. Wer sie aktiv nutzen will, muss selbst prüfen, hochladen, bewerten, ein Aufwand, den Nutzer selten betreiben.

Darüber hinaus ist Erkennung nicht dasselbe wie Abwehr. Selbst wenn ein System verlässlich feststellen könnte, dass ein Text, Bild oder Video von einer KI erstellt wurde, sagt das noch nichts über seine Absicht, seinen Wahrheitsgehalt oder seine Wirkung aus. Für die Unterscheidung von Fake und Fakt mag das in einzelnen Fällen hilfreich sein, doch für die alltägliche Informationsflut ist es kaum ein Schutz.

Und schließlich stellt sich die grundlegende Frage: Was bedeutet es überhaupt noch, wenn in Zukunft, die Mehrheit aller Online-Inhalte ohnehin von KI-Systemen stammen? Jede Kennzeichnung verliert ihren Sinn.

Lösungen, Gegenmaßnahmen und Risikominimierung

Die Probleme, Risiken und Herausforderungen von KI, aber auch KI-Desinformation und Manipulation zu beschreiben, ist vergleichsweise leicht. Lösungen, Abwehr- und Gegenmaßnahmen und Risikominimierungsstrategien zu beschreiben, ist hingegen kompliziert und eine einfache Lösung gibt es nicht. Bereits in unserer vorherigen Publikation zu Desinformation im Allgemeinen haben wir verschiedene Arten von Gegenmaßnahmen, Strategien und Aktionen gegen Desinformation (insgesamt über 20 Einzelpunkte) vorgestellt. Viele davon gelten grundsätzlich auch für KI-Desinformation und Manipulation. Im Folgenden soll es deshalb nur um KI-spezifische Lösungen und Gegenmaßnahmen gehen.

Ähnlich wie bei Maßnahmen gegen Desinformation, Manipulation und Einflussnahme generell, müssen Gegenmaßnahmen in verschiedene Gruppen unterteilt werden.

An erster Stelle kommt *KI-Bildung, Aufklärung und Resilienzbildung* (also der Widerstandsfähigkeit auf Seiten von Nutzern und Publikum) eine zentrale Rolle zu. Programme zur Förderung allgemeiner KI-Kompetenz und KI-spezifischer Medienbildung sind Voraussetzung, um Nutzer in die Lage zu versetzen, KI-generierte Inhalte zu erkennen. Wie angesichts der hier beschriebenen Manipulationsmöglichkeiten deutlich wird, gehören digitale Fähigkeiten und Kenntnisse zum Grundwissen des KI-Zeitalters. Öffentlichkeitskampagnen können das Bewusstsein für die Existenz und Wirkung von KI-Manipulationen schärfen. Darüber hinaus bieten Vorbeugung gegen Täuschungsstrategien durch frühzeitige Aufklärung wirksame Ansätze.

Technologische Gegenmaßnahmen gegen KI-Manipulationen sind deshalb noch wichtiger als zuvor. Solche Gegenmaßnahmen umfassen den Einsatz von KI selbst. KI-basierte Systeme zur Erkennung, Klassifizierung und Moderation von online Inhalten auf Plattformen, Webseiten oder Foren bilden eine der zentralen Verteidigungslinien gegen manipulierte oder synthetische Inhalte. Diese Systeme können Texte, Bilder, Audio- und Videodateien automatisiert analysieren und Anomalien, Manipulationsmuster oder künstlich erzeugte Merkmale identifizieren.

Ergänzend können Verfahren der Kennzeichnung, Nachvollziehbarkeit und Authentifizierung von Inhalten eingesetzt werden. Ziel ist es, die Herkunft, Echtheit und mögliche Manipulation eines online Inhalts maschinenlesbar und überprüfbar zu machen. Schutzsoftware, die KI-generierte Manipulationen erkennt oder abwehrt, stellt eine weitere technologische Ebene dar, die in Kommunikations- und Sicherheitssysteme integriert werden kann.

Rechtliche Gegenmaßnahmen konzentrieren sich auf die Regulierung von Anwendungen und Einsatzbereichen. Dazu gehören Verbote bestimmter KI-Anwendungen, insbesondere solcher, die zur gezielten Täuschung, politischen Beeinflussung, zur Gefährdung der öffentlichen Sicherheit oder zur Erstellung von Pornographie eingesetzt werden können. Ebenso wichtig ist die Regulierung des KI-Einsatzes in politisch, sicherheitsrelevant oder gesellschaftlich sensiblen Bereichen. Vorgeschrieben werden können auch verpflichtende Sicherheitsfunktionen und Schutzmechanismen innerhalb von KI-Systemen, die die Erzeugung von Desinformation technisch erschweren und verringern. Auch algorithmische Transparenz sowie die Regulierung von Empfehlungssystemen und Werbealgorithmen sind zentrale Elemente, um manipulative Einflussnahmen einzudämmen. Ergänzend kommen Haftungsmechanismen hinzu, die auf die Einhaltung dieser Vorschriften abzielen.

Ethische Normen und Richtlinien bilden einen weiteren Pfeiler einer Abwehrstrategie. Sie betreffen vor allem die

KI-generiertes Bild (ChatGPT)

In einer zunehmend automatisierten Welt: Humanoide Roboter übernehmen die Arbeitsplätze, während der Mensch nur noch als Einzelner im Hintergrund steht.

Entwicklung und Anwendung von KI-Systemen selbst. Notwendig sind internationale Produkt- und Sicherheitsstandards, die verbindliche Anforderungen an Sicherheit, Qualität und Nachvollziehbarkeit festlegen. Offizielle Leitlinien zum verantwortungsvollen Umgang mit KI sollten sowohl für Hersteller als auch für Anwender gelten.

Auch *Maßnahmen der Informationssicherheit* gewinnen im Kontext von KI eine neue Bedeutung. Dazu gehören die systematische Speicherung und Überprüfung von Prompts, um Missbrauch, Manipulation oder unerwünschte Ergebnisse nachverfolgen zu können.

Abschließend ist zu betonen, dass die Herausforderungen nicht allein in der Abwehr von KI-Manipulation liegen, sondern

auch im Umgang mit einer neuen Informationsrealität. Die KI-Ära ist durch eine veränderte Informationsumgebung gekennzeichnet: Synthetische Inhalte, automatisiert erzeugte Daten und maschinell generierte Narrative werden in den nächsten Jahren zum Normalzustand. Die Unterscheidung zwischen authentischer und künstlicher Information verliert an Eindeutigkeit, während zugleich die Menge und Geschwindigkeit der Informationsproduktion exponentiell zunehmen. In dieser neuen Umgebung werden Fähigkeiten zur Navigation, zum Management und zur Bewertung von Informationen zu zentralen Kompetenzen. Sie bilden eine neue Form digitaler Medien- und KI-Kompetenz. Systeme zur Klassifikation und Kennzeichnung, vergleichbar mit »Nährwertangaben« für Informationen, könnten künftig dazu beitragen, die Herkunft, Qualität und Zuverlässigkeit digitaler Inhalte transparenter zu machen. Entscheidend wird sein, nicht nur die Risiken synthetischer Information zu kontrollieren, sondern den konstruktiven, verantwortungsvollen und sicheren Umgang mit ihr zu erlernen.

KI und Hacking und Cybersecurity

Die Verbindung von KI und Cybersicherheit betrifft längst nicht mehr nur technische Schutzsysteme oder klassische Hacking-Angriffe, sondern berührt zentrale Fragen der gesellschaftlichen Stabilität. Denn KI ist nicht nur ein Werkzeug zur Analyse oder Automatisierung, sondern ein System, das einerseits auf Daten aufbaut und andererseits aus dieser Datenbasis neue Daten und Informationen generiert. An beiden Enden dieser Prozesse können Daten und Informationen angegriffen und manipuliert werden.

Wenn verfälschte oder absichtlich irreführende Inhalte in Trainingsdaten eingespeist werden, spricht man von Vergiften von Daten. Dabei handelt es sich um gezielte Angriffe auf die Grundlage von KI-Modellen, bei denen manipulierte Informationen verwendet werden, um langfristig die Ausgaben oder Empfehlungen des Systems zu verändern. Das bedeutet: Schon bevor ein KI-Modell Inhalte ausgibt, kann es durch unsichtbare Eingriffe »vergiftet« sein, die Manipulation beginnt also nicht erst am Ausgangspunkt, sondern bereits in tief in der Grundlage des Systems. Die Nutzer der Endanwendung bekommen davon in aller Regeln wenig bis gar nichts mit und haben nur das vom System produzierte Endergebnis als Anhaltspunkt.

Ein anderes Cybersecurity-Beispiel für manipulative Angriffe auf KI-Systeme ist die *Indirect Prompt Injection*. Hier werden scheinbar harmlose Texte, Emails, Anhänge oder Webseiten mit versteckten Anweisungen versehen, die von Sprachmodellen wie Chatbots oder KI-Assistenten automatisch ausgeführt werden. Diese unsichtbaren Befehle, eingebettet in E-Mails, Produktbeschreibungen oder HTML-Code, können dazu führen, dass das KI-System manipulierte Antworten erzeugt, oft

ohne dass Nutzer oder Entwickler dies sofort bemerken. Cybersecurity-Forscher schafften es so z. B. in Email-Anhängen versteckte Befehle an Microsoft Copilot zu senden und dadurch den Output zu manipulieren. Das besonders Frappierende daran war, dass dies einerseits durch einfache Text-Befehle in Email-Anhängen möglich war und dass diese Anhänge andererseits nicht vom Nutzer geöffnet werden mussten, sondern Copilot sie automatisch auslas. Hier reicht es, dass ein KI-gestütztes System automatisch mit manipulierten Inhalten konfrontiert wird. Solche Angriffe lassen sich mit generativer KI vorbereiten und automatisieren, was ihre Verbreitung deutlich vereinfacht.

Ein weiterer Angriffsvektor ist das sogenannte *LLM-Hijacking*, also die gezielte Übernahme oder Umsteuerung großer Sprachmodelle. Dies kann durch die Einbettung von toxischen Inhalten in Feedback- oder Trainingsphasen der Modelle passieren. Besonders bedenklich wird dies, wenn die betroffenen Systeme in sensible Anwendungen integriert sind, etwa in Behörden, Banken, Versicherungen, Nachrichtenportale oder öffentliche Suchsysteme.

Tatsächlich zeigen konkrete Fälle, dass generative KI bereits für klassische Cyberangriffe eingesetzt wird. Russische Hackergruppen nutzten etwa im Rahmen von Angriffen auf die Ukraine KI-generierte Phishing-E-Mails, die aufgrund ihres realistischen Stils besonders glaubwürdig wirkten. Diese Mails enthielten Links oder Anhänge, die weitere KI-generierte Schadsoftware automatisch installierten und dabei von einem KI-Chatbot angeleitet und gesteuert wurden. Nach der Erstinfektion kam es zur systematischen Spionage (*Computer Network Exploitation*), in der sensible Daten abgegriffen wurden. Anschließend verbreitete sich der Schadcode automatisiert weiter und erreichte neue Ziele, ohne dass Benutzer aktiv etwas tun mussten.

KI ist auch in der Cybersecurity in absoluter game changer mit enormen Auswirkungen. Die Verknüpfung und Verwendung von Daten für KI-Training sowie die Integration von KI in alle

möglichen Anwendungen (z. B. Copilot, Handys, Social Media, Autos) bietet eine sehr breite Angriffsoberfläche für Manipulationen. Manipulierte Daten können nun sowohl viel leichter in Systeme rgelangen und auch wieder an den Endnutzer kommen. Dies bedeutet für moderne Informationsgesellschaften ein gesteigertes Manipulationsrisiko.

Gleichzeitig zeigt sich aber auch, dass KI auf der Verteidigungsseite für Cybersecurity eine immer wichtigere Rolle spielt. Sie kann dabei helfen, große Datenmengen in Echtzeit auszuwerten, ungewöhnliche Muster und Verhaltensweisen zu identifizieren (Anomalieerkennung) und damit Hinweise auf Angriffe frühzeitig zu liefern. KI-gestützte Systeme können außerdem automatisch Schwachstellen in Software testen, um Angriffsflächen zu reduzieren, und im Ernstfall auch automatisiert Gegenmaßnahmen einleiten. Gerade in komplexen, vielschichtigen IT-Systemen stellt dies einen wichtigen Fortschritt dar, da menschliche Analysten die Geschwindigkeit und Vielfalt der Angriffe allein nicht mehr bewältigen könnten. Damit steht KI zugleich auf beiden Seiten: Als Werkzeug der Angreifer, um Angriffe realistischer, schneller und schwerer erkennbar zu machen und als Werkzeug der Verteidiger, um Manipulationen früher zu erkennen und gezielter zu reagieren.

KI-generiertes Bild (ChatGPT)

Das Bild stellt die Frage, ob die Menschheit einer rosigen Zukunft entgegensehen wird.

Fazit und Ausblick

Mit jeder neuen Modellgeneration steigen die Möglichkeiten, Inhalte automatisch zu erzeugen, zu verbreiten und zu analysieren. Die nächste Phase ist geprägt durch den Einsatz von KI-Agenten, also Systemen, die nicht nur Informationen verarbeiten, sondern selbstständig Aufgaben ausführen, Entscheidungen vorbereiten oder digitale Prozesse übernehmen. Gleichzeitig schreitet auch die Automatisierung in der Cyberabwehr voran. KI wird zunehmend dazu eingesetzt, Bedrohungen zu erkennen, auf Angriffe zu reagieren und Sicherheitssysteme in Echtzeit anzupassen. Damit entsteht ein Wettrennen zwischen automatisierter Manipulation und automatisierter Verteidigung.

Eine zentrale Herausforderung liegt in der Geschwindigkeit dieser Entwicklung.

KI dringt in immer neue politische und gesellschaftliche Felder vor: KI-Influencer, KI-gepowerte politische Kommunikation, Politiker-Avatare und sogar erste KI-Minister nehmen immer weiter zu. Dies hat vielfältige Auswirkungen auf die moderne Demokratie: Digitale Politiker, datengesteuerte Programme oder algorithmische Verwaltungsakte lösen Ängste vor einer digital simulierten Demokratie aus. Immer mehr politische Entscheidungen beruhen auf Datenanalysen, die durch KI vorverarbeitet werden, eine Entwicklung, die zur Frage führt, ob sich Demokratie schleichend in eine datenbasierte Steuerung transformiert.

Mit der Integration von KI in physische Systeme verschwimmen in der Zukunft zudem die Grenzen zwischen Software und Hardware. Die Verschmelzung von generativer KI mit Geräten des Alltags, etwa in Smartphones, Haushaltsgeräten oder vernetzten Arbeitsumgebungen, führt dazu, dass KI unbemerkt

KI-generiertes Bild (ChatGPT)

Die Journalistin wird von einer unkontrollierbaren Datenwelle aus Fake-News überschwemmt.

Teil fast aller Lebensbereiche wird. Sie wird zunehmend in Anwendungen eingebettet, die zuvor keine Verbindung zu automatisierter Inhaltserzeugung hatten: von Textverarbeitung über Kundenservice bis hin zur Gestaltung von Lern- und Gesundheitssystemen.

Parallel dazu entstehen eigene Plattformen, Informationssysteme und technische Infrastrukturen für KI-generierte Inhalte. Diese Ökosysteme, wie z. B. OpenAIs Sora2-Plattform, funktionieren nicht nur unabhängig von klassischen Medien oder Institutionen, sondern kreisen auch ausschließlich um KI-generierte, also synthetische Inhalte. Die Frage der Unterscheidbarkeit und Erkennung von KI-Inhalten wird in diesen Informationsräumen irrelevant. Dies birgt zahlreiche Herausforderungen, vor allem bei Manipulationen: Die Frage, was im KI-Zeitalter echt, real, faktisch und authentisch ist.

Was bedeutet das alles für einen jeden von uns? KI verändert die Welt – die Arbeitswelt, die Medienwelt, die Welt der

Kommunikation, Cybersecurity, Desinformation und Kommunikation. Aber wir sind KI-Manipulationen und böswilliger KI-Nutzung nicht einfach so ausgeliefert, noch würde es helfen, den Kopf in den Sand zu stecken und zu hoffen, die KI-Revolution werde schon irgendwie an uns vorüberziehen. Eines der wirksamsten Gegenmittel gegen übertriebene Ängste (und Erwartungen), aber auch gegen KI-Manipulationen, ist die regelmäßige, verantwortungsvolle und bewusste Nutzung von KI. Sie ist die beste Sensibilisierung für Stärken und Schwächen von KI, für seltsame Ergebnisse und ein Augenöffner dafür, wie andere sie für Manipulationen einsetzen können. Und KI wird – so oder so – in naher Zukunft in immer mehr andere Technologien (vom Auto bis zum Headset) integriert sein. Wie bei anderen Technologien auch, eröffnet dies viele Möglichkeiten für Manipulation. Es eröffnet aber auch Chancen und die Möglichkeit, Risiken so früh wie möglich zu erkennen und zu mildern. Und auch bei KI gilt: Gerade wegen ihrer Möglichkeiten für böswilligen und manipulativen Einsatz sollte das Feld nicht den Angreifern überlassen werden.

Literatur (Auswahl)

Karen Allen/Christopher Nehring: AI Generated Disinformation in Europe and Africa, hg.: Medienprogramm Subsahara-Afrika der Konrad-Adenauer-Stiftung, 2025 *(https://www.kas.de/en/web/medien-afrika/einzeltitel/detail/-/content/new-study-ai-generated-disinformation-in-europe-and-africa)*

Emilio Ferrara: GenAI Against Humanity: Nefarious Applications of Generative Artificial Intelligence and Large Language Models, in: Journal of Computational Social Science, 2024 *(https://doi.org/10.1007/s42001-024-00250-1)*

Philipp Hacker / Atoosa Kasirzadeh / Lilian Edwards: AI, Digital Platforms, and the New Systemic Risk, in: Computers & Society, 2025 *(https://doi.org/10.48550/arXiv.2509.17878)*

Katja Munoz: »Systematische Manipulation sozialer Medien im Zeitalter der KI – Eine wachsende Bedrohung für die demokratische Meinungsbildung«, hg.: DGAP *(https://dgap.org/de/forschung/publikationen/systematische-manipulation-sozialer-medien-im-zeitalter-der-ki-0)*

Christopher Nehring: »Information apocalypse or overblown fears—what AI mis and disinformation is all about? Shifting away from technology toward human reactions«, in: Politics & Policy, 2024 (https://onlinelibrary.wiley.com/doi/full/10.1111/polp.12617)

Christopher Nehring: »On the way to deep fake democracy? Deep fakes in election campaigns in 2023«, in: European Political Science, 23(4), 454–473, 2024 *(https://www.cambridge.org/core/services/aop-cambridge-core/content/view/8F97B6AD4C40B195B369696926B5F7EB/S168043330001251Xa.pdf)*

Maria Pawelec: Wenn der Schein trügt – Deepfakes und die politische Realität. Politische Manipulation und Desinformation, hg: Bundeszentrale für Politische Bildung, 2024 *(https://www.bpb.de/lernen/bewegtbild-und-politische-bildung/556305/politische-manipulation-und-desinformation/)*

Nina Schick: Deep Fakes and the Infocalypse, Ottawa, 2020.

tlpb:
Thüringer
Landeszentrale für
politische Bildung

Künstliche Intelligenz (KI) ist seit Ende 2022 ein Hype- und Modethema geworden, das die schlimmsten Ängste und kühnsten Hoffnung zugleich anspricht. Daher ist der Bedarf an unaufgeregtem Orientierungswissen zu diesem Thema groß. Dies ist das Ziel der hier vorliegenden Publikation: Eine der größten Ängste und Risiken im Zusammenhang mit KI aufgreifen, erklären, in ihren Kontext einordnen und Lösungsansätze vorzustellen. So sollen (unbegründete) Ängste abgebaut, KI-Wissen und -Bildung verbreitet und eine sicherer(e) (»safe & secure AI«) und vertrauenswürdigere (»trustworthy AI«) KI vorangetrieben werden.

ISBN: 978-3-910740-71-6