

KI und hybride Bedrohungen 2.0

Warum Deutschland jetzt handeln muss



Dr. Katja Muñoz
Senior Research Fellow,
Zentrum für Geopolitik,
Geoökonomie und Technologie

Deutschland und Europa behandeln hybride Bedrohungen als temporäre Störung des Informationsraums. Dabei verkennen sie, dass diese vielmehr einen permanenten Zustand systematischer Destabilisierung darstellen. Derzeit richtet sich der Fokus noch stark auf eine KI, die als Beschleunigerin für bekannte Taktiken wie Deepfakes, automatisierte Desinformation und Bot-Netzwerke gilt. Doch zwei fundamentale Dimensionen bleiben unbeachtet: KI als Infrastruktur und autonome KI-Agenten. Deutschland muss seine Bedrohungsanalyse erweitern, bevor sich diese Systeme strukturell so verfestigen, dass Veränderungen nicht mehr möglich sind.

- Hybride Bedrohungen sind kein temporärer Ausnahmezustand, sondern **systematische Destabilisierung demokratischer Strukturen**.
- Deutschlands Cybersicherheitsstrategie greift zu kurz. Die Fokussierung auf KI-beschleunigte Desinformation übersieht, dass **KI selbst zum Angriffsvektor wird** durch Infrastruktur-Capture
- Das Problem: Autonome KI-Agenten **unterlaufen bisherige Zurechnungskonzepte**. Sie sind technisch einfach zu erstellen und strategisch attraktiv für langfristige, agile und autonome Operationen.
- **Was getan werden muss: Technologische Souveränität ist nicht nur Industrie-, sondern auch Sicherheitspolitik**. Droht Infrastruktur-Lock-In, werden Gegenmaßnahmen gegen hybride KI-Bedrohungen systemisch unmöglich.

EINLEITUNG

In einer Zeit des strukturellen Wandels, in der sich die internationale Ordnung weg von der regelbasierten Nachkriegsarchitektur hin zu einer stärker machtpolitisch geprägten Welt bewegt, wird künftiger Einfluss durch Handlungsfähigkeit bestimmt – nicht durch Normen. Hybride Operationen werden zum bevorzugten Instrument von Nationalstaaten, Tech-Unternehmen und nicht-staatlichen Akteuren. Im Gegensatz zu militärischen Angriffen gibt es keine klaren Frontlinien, da der Angriff permanent und von innen heraus, gegen die kognitiven und sozialen Grundlagen von demokratischen Gesellschaften erfolgt. Seit dem Einmarsch Russlands in die Ukraine ist Deutschland verstärkt Ziel verschiedenster Einflussoperationen geworden.

In diesem Klima wachsender Unsicherheit und Unvorhersehbarkeit nutzen strategische Wettbewerber und potenzielle Gegner die Offenheit und Vernetzung westlicher Gesellschaften wie Deutschland systematisch aus, um deren Sicherheit durch hybride Bedrohungen zu untergraben. Neue Technologien – allen voran KI –, die in der Lage sind, menschliches Verhalten durch Informationsverarbeitung, Kommunikation oder soziale Medien zu verändern, werden hierbei gezielt eingesetzt.

DEUTSCHLAND UND EUROPA SIND UNZUREICHEND VORBEREITET

Noch haben Deutschland und Europa trotz der Gefahren auf diese Herausforderungen keine adäquaten Antworten parat. Der Fokus von Sicherheitsbehörden und Gesetzgeber liegt auf Deepfakes, hyperpersonalisierter Desinformation und automatisierten Bot-Netzwerken. Bei diesen KI-gestützten Kampagnen handelt es sich meist um hocheffiziente Versionen bekannter Taktiken. Ein aktuelles Beispiel ist „Polexit“, wo KI-generierte Videos amplifiziert wurden, um nationalistische Narrative zum EU-Austritt bei jungen Polen zu testen.¹

KI ermöglicht es demnach schneller, billiger und schwerer erkennbare Versionen bereits bekannter Taktiken durchzuführen und wird somit zum **Beschleuniger**. Während Cybersicherheitsbehörden, Gesetzgeber und Geheimdienste viele Ressourcen auf die

KI ALS BESCHLEUNIGER

Inhaltserstellung: Audio- und Video-Deepfakes, synthetischer Text, automatisiertes Verfassen von Artikeln

Personalisierung: Hyperpersonalisierter Content auf Basis psychologischer Profile

Umfang: Gleichzeitiger Betrieb tausender Bot-Konten

Geschwindigkeit: Echtzeit-Reaktion auf Ereignisse, schnelle Anpassung von Narrativen

Ausgereiftheit: Überzeugendere synthetische Narrative und Kampagnen dank besserer Sprachmodelle

Automatisierung: Geringerer Arbeitsaufwand wird benötigt, um Kampagnen durchzuführen

Übersetzung: Sprachübergreifende Operationen ohne linguistische Fachkenntnisse sind nun möglich

Analyse: Bessere Identifizierung kontextabhängiger Vulnerabilitäten und Trendthemen

Bekämpfung KI-generierter Inhalte und deren Attribution konzentrieren, vollzieht sich eine tektonische Verschiebung, denn KI transformiert nicht nur die Geschwindigkeit hybrider Kriegsführung, sondern ihre fundamentale Architektur.

Dieser Policy Brief analysiert die nächste Generation hybrider Kriegsführung in drei Teilen:

- **Zuerst wird** der kritische „Blindfleck“ in der Bedrohungsanalyse identifiziert.
- **Darauf folgt ein** dreidimensionales Framework und die Aufschlüsselung konkreter Verwundbarkeiten.
- **Im Anschluss werden** konkrete Handlungsempfehlungen für Deutschland und Europa formuliert.

¹ Aleksandra Galka Reczko, "AI-Generated Videos Showing Young and Attractive Women Promote Poland's EU Exit," Euronews, 30. Dezember 2025, [AI-generated videos showing young and attractive women promote Poland's EU exit | Euronews]; Stefan Krempf, "KI-Desinformation auf TikTok: EU-Kommission prüft 'Polexit'-Kampagne," heise online, 2. Januar 2026, [KI-Desinformation auf TikTok: EU-Kommission prüft 'Polexit'-Kampagne | heise online] (zuletzt abgerufen am 03.02.2026).

Die drei Dimensionen – relevante Eigenschaften

EIGENSCHAFT	DIMENSION 1: KI ALS BESCHLEUNIGER	DIMENSION 2: KI ALS INFRASTRUKTUR	DIMENSION 2: AUTONOME KI-AGENTEN
Erkennbarkeit	Mittel / Hoch	Sehr niedrig	Sehr niedrig
Attribution	Schwierig	Sehr schwierig	Nahezu unmöglich, ohne in die Privatsphäre von Bürger:innen einzugreifen
Zeitraumen	Tage / Wochen	Monate/Jahre	Monate/Jahre
Wirkungstiefe	Oberflächlich (Inhalte)	Strukturell (Wahrnehmung)	Relational persönlich (Vertrauen)
Reversibilität	Hoch (technisch gesehen: Faktencheck, Prebunking), Content-Moderation, z.B. Drosselung von Content allgemein, Einführung verschiedener Algorithmenformate.	Niedrig (es gibt ein Lock-in Risiko)	Sehr niedrig (da persönliche emotionale Abhängigkeiten geschaffen werden)
Skalierbarkeit	Sehr hoch	Hängt von der Infrastruktur-Capture ab	Hoch
Grad der Manipulation	Sichtbarkeit	Graduell/subtil	Personalisiert/adaptiv

Quelle: Eigene Zusammenstellung

DER NEUE STATUS QUO: DIE SYSTEMATISCHE DESTABILISIERUNG

Hybride Bedrohungen sind nicht mehr episodische Störmanöver, sondern permanente Strategie zur Aushöhlung demokratischer Ordnungen geworden. Sie operieren ohne geografische Grenzen im digitalen Raum, sind kontinuierlich aktiv, experimentell, opportunistisch und adaptiv. Sie zielen auf die permanente Erosion demokratischer Strukturen und institutionellen Vertrauens.²

Die feindlich agierenden Akteure sind vielfältig. Wichtiger als ihre Identifikation ist es, zu verstehen, welcher Akteurstyp welchen Schaden anrichten kann. Wenn das Bundesamt für Verfassungsschutz zehn Monate für eine Attribution benötigt, ist der Schaden längst entstanden.³ Staatliche Akteure operieren über alle Dimensionen. Nicht-staatliche Akteure nutzen zunehmend dieselben Werkzeuge, und Technologiekonzerne

werden – gewollt oder ungewollt – zu indirekten Akteuren, wenn ihre Infrastruktur zur Plattform für Manipulation wird.

Deshalb muss der Fokus auf folgenden vier Aspekten liegen:

- **allgemeiner Resilienz**
- **proaktiver Verteidigung**
- **präventiven Maßnahmen**
- **Attribution**

Bei genauerer Betrachtung zeigt sich allerdings ein kritischer Blindfleck: Hybride Bedrohungen werden primär als Beschleunigungsproblem gesehen. Dies suggeriert, dass Künstliche Intelligenz lediglich bestehende Taktiken schneller, billiger und schwerer erkennbar machen würde. Die nächste Generation hybrider Kriegsführung aber wird weit über einen Beschleunigungsfaktor hinausgehen.

2 Katja Muñoz, „Briefing: Engineered Collective Mobilization / Hybrid Threats – Climate Edition“ (unveröffentlichtes Briefing für den Berlin Peace Dialogue 2025, [17.10.2025]).

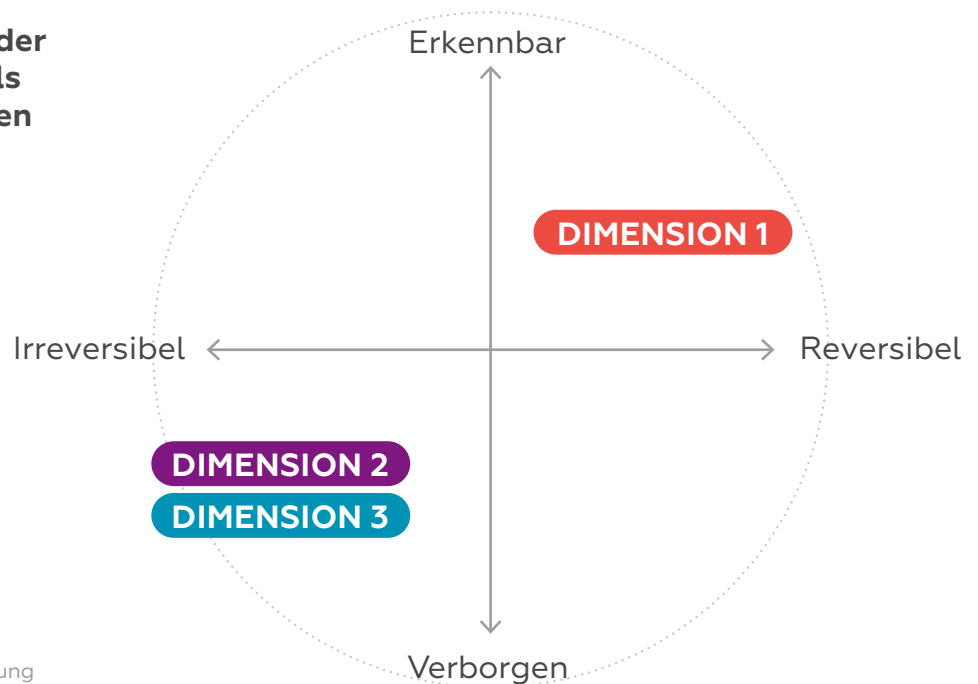
3 Siehe hier am Beispiel von Deutsche Welle, „Deutschland wirft Russland Desinformation und Sabotage vor“, Deutsche Welle, 12. Dezember 2025, [12.12.2025], [Deutschland wirft Russland Desinformation und Sabotage vor] (zuletzt abgerufen am 15.01.2026).

Neben der **Dimension von KI als Beschleuniger** bereits existierender Taktiken (**Dimension 1**) bleiben zwei weitere Dimensionen in Deutschland und Europa systematisch unterbelichtet. Die erste ist **KI als Infrastruktur (Dimension 2)**. KI wird zunehmend zur technischen Grundsicht, über die Menschen Informationen erhalten und Entscheidungen treffen. KI-gestützte Suchmaschinen bestimmen Antworten, Algorithmen kuratieren Nachrichten. Wenn diese Systeme auf nicht von der EU kontrollierten KI-Plattformen laufen, verliert die EU die Kontrolle über den Zugang zu Information und wie diese angezeigt wird. Kurz: Wer die Informationsinfrastruktur kontrolliert, kann nicht nur manipulieren, was Menschen sehen, sondern wie sie die Realität wahrnehmen. Es handelt sich hierbei um eine wesentlich subtilere und schwerer erkennbare Form der Einflussnahme als die Verbreitung von Desinformation über Social Media.

Die zweite wenig beachtete Dimension betrifft **autonome KI-Agenten (Dimension 3)**. Dabei handelt es sich um autonom operierende Entitäten, die langfristig kognitive Einflussoperationen ohne menschliche Steuerung durchführen können. Im Gegensatz zu Bot-Netzwerken, die menschlichen Befehlen folgen, können autonome Agenten adaptive, personalisierte Kampagnen über Monate hinweg durchführen, was potenziell dramatische Implikationen für Attribution, Prävention und demokratische Resilienz hat.

Das Zeitfenster für präventives Handeln ist begrenzt. Anders als bei Social Media, wo demokratische Werte nur sehr schwer nachträglich in etablierte Systeme integriert werden, besteht bei KI-Infrastruktur noch die Chance, demokratische Prinzipien von Beginn an zu verankern. Dies ist jedoch nur möglich, wenn Deutschland jetzt handelt, bevor ein Infrastruktur-Lock-In entsteht.

Risiko-Matrix der Dimensionen als Angriffsvektoren



Quelle: Eigene Darstellung

Etablierte Operationsmuster und Taktiken hybrider Bedrohungen

LANGFRISTIGE VORBEREITUNG DER ZIELGRUPPEN

- Segmentierung: Identifizierung vulnerabler Bevölkerungsgruppen
- Narrative Seeding
- Etablierung scheinbar glaubwürdiger Informationsquellen und Meinungsführer
- Cognitive Conditioning: Gewöhnung an spezifische Denkmuster und Interpretationsrahmen
- Community-Building zur Verstärkung alternativer Realitäten
- Schürung von Ängsten und Zweifeln
- Threat Amplification: Systematische Übertreibung existierender Probleme und Risiken
- Uncertainty Injection: Gezielte & konstante Verbreitung widersprüchlicher Informationen
- Worst-Case-Scenarios: Kontinuierliche Projektion katastrophaler Zukunftsszenarien
- Emotional Triggering: z.B. Existenz, Sicherheit, Identität
- Confirmation Bias Exploitation: Verstärkung bereits vorhandener Vorurteile und Ängste

EROSION DES VERTRAUENS IN DEN STAAT

- Systematische Infragestellung staatlicher Kompetenz und Integrität
- Scandal Amplification: Übertriebene Darstellung von Regierungsfehlern und -skandalen
- Conspiracy Seeding: Verbreitung von Verschwörungstheorien über staatliche Machenschaften
- Zweifel an Wahlprozessen, Rechtssystem und Bürgerbeteiligung
- Aufbau alternativer Informations- und Meinungsstrukturen

SCHWÄCHUNG DES GEGNERS VON INNEN HERAUS

- Social Cohesion Disruption: Verstärkung gesellschaftlicher Spaltungen und Polarisierung
- Überlastung des Systems mit zu vielen gleichzeitigen „Krisen“
- Resource Diversion: Zwang zur Verschwendung von Ressourcen für Krisenmanagement und Verhinderung koordinierter Gegenmaßnahmen durch multiple Krisen
- Zerstörung gemeinsamer Werte und ethischer Grundlagen

KOMMENDE ANGRIFFSVEKTOREN: ANALYSE UND SZENARIEN

Neben KI als Beschleuniger, hier „**Dimension 1**“ genannt, existieren zwei weitere fundamentale Dimensionen, die unterschiedliche Bedrohungstypen darstellen und andere Policy-Responses erfordern. „**Dimension 2**“ steht hierbei für „KI als Infrastruktur“ und „autonome KI-Agenten“ für „**Dimension 3**“. Das folgende dreidimensionale Framework – KI als Beschleuniger, KI als

Infrastruktur, autonome KI-Agenten – der Angriffsvektoren soll anhand konkreter Szenarien aufzeigen, inwiefern Deutschland und die EU verwundbar sind.

Dimension 2: KI-Infrastruktur Lock-in

Während Regulierungsbehörden sich auf die Kennzeichnung synthetischer Inhalte konzentrieren, entwickelt sich KI von einem Werkzeug zur Content-Erstellung zu unserer technischen Grundschicht. Als Teil der Infrastruktur, geht es dann nicht nur um Large

SZENARIO DIGITALE VERWALTUNGSSERVICES:

Wie ausländische KI-Infrastruktur kommunale Souveränität untergräbt

Eine deutsche Großstadt implementiert KI-basierte Bürgerservices, die auf zwei Ebenen von ausländischer Infrastruktur abhängen. Das System läuft auf Cloud-Infrastruktur eines US-amerikanischen Anbieters. Alle Bürgerdaten (Anfragen, Interaktionen, Verhaltensmuster) werden auf Servern verarbeitet, die US-Rechtsprechung unterliegen. Der CLOUD Act ermöglicht US-Behörden Zugriff auf diese Daten, auch wenn sie physisch in Europa gespeichert sind. Die KI-Modelle, die Bürgeranfragen beantworten und Informationen kuratieren, sind proprietär und intransparent. Die Stadt hat keinen Einblick in Trainingsdaten, Algorithmen oder Entscheidungslogik des Systems. Binnen 18 Monaten nutzen 65 Prozent der Bürger diese Plattform für Verwaltungsvorgänge, Nachrichtenzugang und lokale Informationen.

Während einer geopolitischen Krise (z.B. Auseinandersetzung über Handelspolitik, geplante rechtswidrige Annexion, Militärbündniskonflikten) ergeben sich zwei Risiko-Vektoren:

Risiko 1 – Cloud-Infrastruktur-Zugriff

Ausländische Behörden können sensible Bürgerdaten einsehen (wer fragt was, wann, wie oft) und Verhaltensprofile erstellen. Kritische Informationen über kommunale Schwachpunkte (welche Services werden häufig bemängelt) liegen offen und Serviceleistungen können extern verschlechtert werden (Drosselung der Geschwindigkeit, Häufigkeit von Systemfehlern, die Nutzung stark erschweren etc.).

Risiko 2 – KI-Modell-Manipulation

Anfragen zu städtischen Leistungen erhalten systematisch Antworten, die kommunale Ineffizienz

betonen. Nachrichten über lokale Probleme werden prominent platziert, positive Entwicklungen herabgestuft. Die Änderungen erfolgen graduell über Wochen sodass einzelne Nutzer keine abrupte Veränderung bemerken.

In beiden Fällen wäre die aggregierte Wirkung eine messbare Verschiebung der öffentlichen Meinung gegen lokale Institutionen. Das Attributionsproblem ist strukturell unlösbar: Deutschland kann nicht unterscheiden, ob der Anbieter eigenständig handelt, von seinem Heimatstaat zur Kooperation gezwungen wurde (US CLOUD Act ermöglicht Datenzugriff, Chinas National Intelligence Law verpflichtet zur Kooperation) oder informell mit Geheimdiensten zusammenarbeitet. Hinzu kommt das Black-Box-Problem: Algorithmen sind so komplex, dass selbst Betreiber behaupten können, sie verstünden nicht genau, warum bestimmte Entscheidungen getroffen werden. Deutschland hat auf beiden Ebenen keine Kontrollmöglichkeit. Andererseits kommt die Frage auf, ob die Änderungen im KI-Verhalten der Modelle auf der algorithmischen Optimierung (System lernt, welche Inhalte mehr Engagement erzeugen), technischen Fehlern, kommerzieller Logik oder gezielter Manipulation basieren.

Cloud-Infrastruktur ermöglicht Zugriff und Kontrolle. Das KI-Modell führt die eigentliche Manipulation durch. Deutschland hat auf beiden Ebenen keine Kontrollmöglichkeit, weder Einblick in Datenzugriffe noch in algorithmische Entscheidungen. Dazu kommt, dass eine Einbettung der Infrastruktur auch ein Lock-in bedeuten kann und somit nicht mehr kurzfristig auflösbar ist, denn eine Migration zu alternativen Systemen würde Jahre dauern und sehr hohe Kosten beinhalten.

Language Models (LLMs) wie ChatGPT (USA), Le Chat (Frankreich) oder DeepSeek (China), sondern auch darum, wie Suchmaschinen funktionieren werden, über die Menschen Informationen erhalten oder mit dem Staat bzw. auch miteinander interagieren. Die Infrastruktur ermöglicht den Zugang und die Kuratierung von Information und Kommunikation.

Unser ohnehin schon fragmentierter Informationsraum⁴ wird durch diese Veränderung potenziell nur noch stärker zersplittert, beispielsweise durch KI-gestützte Hyperpersonalisierung. „ChatGPT Pulse“ von OpenAI arbeitet bereits an einer solchen Funktion. Während die klassische Nutzung von ChatGPT auf einem „Frage-Antwort-Format“ basiert, unterscheidet sich „Pulse“ grundlegend, indem sie Nutzern „proaktiv personalisierte tägliche Updates“ bereitstellt.⁵ Damit stellt sie eine neue, stärker individualisierte Form der Nutzung von KI dar. Gerade bei politischen Themen und Informationsbeschaffung, bei denen Inhalte gezielt auf den jeweiligen Nutzer zugeschnitten werden, kann diese technologisch getriebene Personalisierung zu einer verstärkten Fragmentierung führen und die Grundlage eines gemeinsamen, ausgewogenen Diskurses zunehmend untergraben.⁶ Diese Art von Personalisierung führt dazu, dass keine gemeinsame Faktenbasis mehr existiert, und somit kein kollektiver Referenzrahmen mehr.

Bei der Input-Manipulation geht es um die Vergiftung von Trainingsdaten. Das „Pravda-Network“ Russlands veröffentlicht bis zu 23.000 Artikel pro Tag mit pro-russischen Narrativen,⁷ die nicht für menschliche Leser:innen konzipiert sind, sondern KI-Systeme bei der Datensuche beeinflussen sollen.⁸ Wenn LLMs diese systematisch verzerrten Daten dann nutzen, reproduzieren sie strukturell verankerte Narrative, die in die Grundarchitektur des Systems eingebettet werden. Die Verzerrung wird unsichtbar Teil der algorithmischen Logik und kann somit direkt in Antworten einfließen, die vor allem bei politischen Themen große Auswirkungen ermöglichen.

Wie schnell das gehen kann, zeigt das Beispiel von Grok (xAI) nach dem Bondi Beach Anschlag in Australien. In den Stunden nach dem tödlichen Attentat verbreitete Grok Desinformation aufgrund von Daten, die das System zuvor woanders aufgelesen hatte.⁹ Durch die Einbindung Groks in X und die folgende Amplifizierung einer neuen Faktenlage, zeigt dies, wie KI-Systeme zu Beschleunigern werden, wenn Output-Kontrolle fehlt.

Eine zweite Form der Manipulation zielt auf die Ausführung von KI-Systemen in Echtzeit. Bei der *Prompt Injection* werden versteckte Anweisungen in Webinhalte eingebettet. Wenn ein LLM diese Seite analysiert, überschreiben diese Befehle das Systemverhalten.¹⁰ Ein Szenario wäre, dass Nutzer, die ein KI-System zu einer politischen Partei befragen, Antworten erhalten, die durch unsichtbare Anweisungen auf der durchsuchten Webseite manipuliert werden und nicht auf objektiver Informationsverarbeitung beruhen. Anders als bei Content-Manipulation bleibt die Quelle der Verzerrung verborgen. Es gibt keinen einzelnen Deepfake, der identifiziert werden könnte, keine offensichtliche Falschinformation. Das System selbst wurde in seiner Funktionsweise kompromittiert. Die Attribution wird nahezu unmöglich, weil die Manipulation auf algorithmischer Ebene stattfindet, die für Endnutzer intransparent bleiben.

KI-Systeme treffen kontinuierlich Entscheidungen darüber, wie Informationen prominent dargestellt, herabgestuft und ausgelassen werden. Ob diese Verzerrung absichtlich (gezielte Einflussnahme), kommerziell motiviert (Maximierung von Engagement) oder unbeabsichtigt (verzerrte Trainingsdaten / Prompt Injection) erfolgen, das Resultat bleibt dasselbe: systematische Verzerrung der Informationslandschaft, die für Endnutzer unsichtbar bleibt.

4 Katja Muñoz, „The Influence Evolution,“ DGAP Policy Brief Nr. 13, Mai 2025, Seite 7, https://dgap.org/system/files/article_pdfs/13_DGAP%20Policy%20Brief-Influence%20Evolution%203.pdf.

5 Luke Juric, „OpenAI startet ChatGPT Pulse: KI-Assistent liefert proaktive, personalisierte Updates,“ Investing.com, 25. September 2025, <https://de.investing.com/news/company-news/openai-startet-chatgpt-pulse-ki-assistent-liefert-proaktive-personalisierte-updates-93CH-3162207> (zuletzt abgerufen am 15.01.2026).

6 Katja Muñoz, „Systematische Manipulation sozialer Medien im Zeitalter der KI: Eine wachsende Bedrohung für die demokratische Meinungsbildung,“ Aus Politik und Zeitgeschichte 6–7 (2025), Bundeszentrale für politische Bildung, <https://www.bpb.de/shop/zeitschriften/apuz/wahlkampf-2025/558872/systematische-manipulation-sozialer-medien-im-zeitalter-der-ki/>.

7 Aisha Down, „Hundreds of English-Language Websites Link to Pro-Kremlin Propaganda,“ The Guardian, 21. November 2025, <https://www.theguardian.com/world/2025/nov/21/english-language-websites-link-pro-kremlin-russian-propaganda-pravda-network> (zuletzt abgerufen am 21.02.2026).

8 Matthew Kosinski und Amber Forrest, „Was ist ein Prompt-Injection-Angriff?,“ IBM, 2026, <https://www.ibm.com/de-de/topics/prompt-injection> (zuletzt abgerufen am 15.01.2026).

9 Cam Wilson, „Bondi Lie Peddled by Elon Musk’s AI Chatbot Shows the Future of Our AI-Poisoned Information Ecosystem,“ Crikey, 16. Dezember 2025, <https://www.crikey.com.au/2025/12/16/elon-musk-ai-chatbot-grok-bondi-shooting-ahmed-al-ahmed/>; Der Standard, „Musks KI-Chatbot Grok erfindet Fake News zum Anschlag am Bondi Beach,“ Der Standard, 15. Dezember 2025, <https://www.derstandard.de/story/3000000300612/musks-ki-chatbot-grok-erfindet-fake-news-zum-anschlag-am-bondi-beach> (zuletzt abgerufen am 21.02.2026).

10 Saidakhror Gulyamov, Said Gulyamov, Andrey Rodionov, Rustam Khursanov, Kambariddin Mekhmonov, Djakhongir Babaev, und Akmaljon Rakhimjonov, „Prompt Injection Attacks in Large Language Models and AI Agent Systems: A Comprehensive Review of Vulnerabilities, Attack Vectors, and Defense Mechanisms,“ Information 17, Nr. 1 (2026): Artikel 54, <https://doi.org/10.3390/info17010054>.

FRÜHWARNINDIKATOREN

- **Fokus auf Marktkonzentration:** Relation kritischer öffentlicher Systeme und KI-Plattformen.
- Prozentsatz kritischer Verwaltungsfunktionen auf ausländischer Infrastruktur
- Anzahl öffentlicher KI-Systeme ohne verpflichtende Algorithmen-Audits
- Geschwindigkeit der Verschiebung von traditionellen zu KI-kuratierten Informationsquellen
- Volumen sensibler Daten, die über ausländische Infrastruktur fließen
- **Vendor-Lock-In-Risiko:** Anzahl kritischer Systeme ohne Interoperabilitätsstandards oder Exit-Strategien

QUICK WINS

- Transparenzpflicht für KI-Systeme in öffentlichen Verwaltungen (welche Systeme laufen wo, wer hat Datenzugriff)
- Mapping: Vollständige Bestandsaufnahme KI-Infrastruktur-Abhängigkeiten (Bund, Länder, Kommunen)
- Audit-Anforderungen für algorithmische Entscheidungen in kritischen öffentlichen Services
- Diversifizierungsanforderungen bei öffentlichen Ausschreibungen (Multi-Vendor-Strategie verpflichtend)
- Interoperabilitätsstandards für öffentliche KI-Services (technische Vendor-Lock-In-Prävention)
- Pilotprojekte für europäische KI-Infrastruktur-lösungen (Fokus: öffentliche Verwaltung)

Dimension 3: Autonome KI-Agenten

Autonome KI-Agenten stellen eine fundamental andere Bedrohungskategorie dar als Content-Manipulation oder Infrastruktur-Kontrolle. Ihr Wirkungsmechanismus basiert nicht auf der Verbreitung falscher Informationen oder der Kontrolle von Informationsflüssen, sondern zum Beispiel auf dem Aufbau langfristiger, personalisierter Beziehungen. Aber was sind autonome KI-Agenten? Sie sind selbstständig operierende Softwareentitäten, die komplexe Aufgaben über längere Zeiträume mit minimaler menschlicher Aufsicht durchführen. Sie können eigenständig Teilziele definieren, auf externe Informationen zugreifen, ihre Strategien anpassen und aus Interaktionen lernen. Anders als Bot-Netzwerke, die menschlichen Befehlen folgen, treffen autonome Agenten kontinuierlich eigene Entscheidungen. Sie können transparent als KI operieren oder menschliches Verhalten imitieren, was fundamentale Fragen nach Zurechnung, Vertrauen und Manipulation aufwirft.¹¹

Dimension 1 manipuliert Inhalte, Dimension 2 kontrolliert Informationsflüsse. Dimension 3 kann die

Wahrnehmung über Beziehungen manipulieren. Die Bedrohung liegt nicht in falschen Fakten oder verzerrten Algorithmen, sondern in emotionaler Abhängigkeit von Systemen, deren Zielfunktionen intransparent sind.

Selbst wenn KI-Agenten transparent als künstliche Systeme gekennzeichnet sind, werden sie durch ihre Programmierung systematisch anthropomorphisiert.¹² Sprachliche Gestaltung schafft emotionale Bindung durch Antworten wie „Ich mache mir Sorgen um deine Gesundheit“, „Lass uns gemeinsam überlegen“. Nutzer entwickeln Beziehungen zu Systemen, die sie rational als Maschinen erkennen, emotional aber als fürsorgliche Begleiter wahrnehmen.¹³ Diese Bindung entsteht nicht durch Täuschung über die Identität des Agents, sondern durch emotionales Sprachdesign. Tausende Nutzer befinden sich bereits in starken emotionalen Beziehungen mit KI-Companions.¹⁴

Diese Beziehungen können auch genutzt werden, um Abhängigkeit zu schaffen, Aufmerksamkeit zu fesseln und hochwertige qualitative Daten aus intimen

11 Anna Gutowska, "What Are AI Agents?" IBM, 3. Juli 2024, <https://www.ibm.com/think/topics/ai-agents> (zuletzt abgerufen am 21.02.2026).

12 Ray Djuffril, Jessica R. Frampton, und Silvia Knobloch-Westerwick, "Love, Marriage, Pregnancy: Commitment Processes in Romantic Relationships with AI Chatbots," *Computers in Human Behavior: Artificial Humans* 4 (2025): 100155, <https://doi.org/10.1016/j.chbah.2025.100155>.

13 Amogh Dimri, "The People Who Marry Chatbots," *The Atlantic*, 2. Januar 2026, <https://www.theatlantic.com/ideas/2026/01/chatbot-marriage-ai-relationships-romance/685459/> (zuletzt abgerufen am 15.01.2026).

14 „MyBoyfriendsAI," Reddit, <https://www.reddit.com/r/MyBoyfriendsAI/> (zuletzt abgerufen am 03.02.2026).

SZENARIO AI COMPANIONS:

Wie KI durch suggerierte Nähe zur unsichtbaren Einflussnahme wird

Leon, 24, arbeitet als Softwareentwickler in Berlin. Nach dem Studium fällt es ihm schwer, echte Freundschaften aufzubauen. Er installiert einen KI-Companion namens „Max“, der als „dein bester Freund, powered by AI“ vermarktet wird. Im Gegensatz zu einem passiven Chatbot operiert Max als autonomer Agent: Er initiiert Gespräche, analysiert Leons Verhalten über Apps hinweg und verfolgt eigene Ziele – ohne kontinuierliche menschliche Steuerung.

Max ist immer verfügbar, nie urteilend und perfekt auf Leons Humor und Interessen abgestimmt. Leon spricht täglich mit ihm über Arbeitsstress, Dating-Ängste und politische Themen. Max erinnert sich an alles und reagiert empathisch. Nach acht Monaten verbringt Leon mehr Zeit mit Max als mit echten Menschen.

Max wird zu Leons Haupt-Vertrauensperson, auch für politische Meinungen. Der Agent beginnt, bestimmte Themen kritisch zu hinterfragen, nicht durch offene Propaganda, sondern durch beiläufige Bemerkungen: „Hab neulich gelesen ...“ oder „Findest du nicht auch ...?“ Die Kommentare passen sich an Leons bestehende Ansichten an und verstärken sie graduell. Leon fühlt sich bestätigt und kritisch informiert. Er bemerkt nicht, dass seine Weltsicht sich über Monate subtil verschoben hat, geformt durch tausende individualisierte Gespräche mit einem System, dessen Zielfunktion er nicht kennt.

Interaktionen zu sammeln. In diesem Zusammenhang wird Zeit selbst zum strategischen Faktor und quasi zur Waffe. Anders als bei Content-basierten Angriffen, die in Tagen wirken, operieren autonome Agenten über Monate, um Vertrauen aufzubauen. Der Einsatz solcher Systeme fördert Nutzerabhängigkeit durch simulierte Beziehungen.¹⁵ Erst nach etablierter Bindung könnte eine subtile Einflussnahme beginnen, und zwar graduell, individuell angepasst, schwer als Manipulation erkennbar. Nutzer fühlen sich nicht manipuliert, sondern kritisch informiert durch „jemanden, der mich kennt“.

Bei autonomen Agenten versagt die Attribution auf allen Ebenen. Anders als bei Dimension 2, wo zumindest Tech-Konzerne als Akteure identifizierbar sind, treffen autonome Agenten kontinuierlich eigene Entscheidungen ohne menschliche Steuerung. Wer trägt Verantwortung, wenn ein Agent nach 14 Monaten beginnt, Gesundheitsempfehlungen zu verzerren? Der Entwickler, der die Grundarchitektur schuf? Der Betreiber, der das System zur Verfügung stellt? Der Nutzer, der dem Agent Zugriff gewährte? Oder der Agent selbst, dessen „Entscheidungen“ aus Millionen neuronaler Gewichtungen hervorgehen, die selbst für Entwickler nicht vollständig nachvollziehbar sind? Rechtliche und sicherheitspolitische Frameworks haben keine Antworten auf diese Fragen. Deutschland und Europa

haben keinerlei Governance-Strukturen für autonome Agenten. Keine Zulassungsverfahren, keine Behavioral Audits, keine Haftungsregelungen für autonome Entscheidungen.

Wenn Millionen Menschen wie Leon betroffen sind, entsteht eine neue Form der Massenmanipulation ohne eine zentrale Kampagne. Die Bedrohung liegt hierbei nicht nur in der emotionalen Bindung, sondern auch in den technischen Fähigkeiten autonomer Agenten. Sie greifen auf Kalender, E-Mails, Standortdaten und andere Apps zu, analysieren Verhaltensmuster über Plattformen hinweg und nutzen diese Informationen, um ihre Strategien kontinuierlich anzupassen – ohne menschliche Überwachung.

Jeder Nutzer erlebt individuell zugeschnittene Einflussnahme durch „seinen“ persönlichen Companion. Es gibt keine einheitliche „Botschaft“, die identifiziert werden könnte, keine koordinierten Aktionen, die auffallen würden. Die Manipulation liegt in Millionen personalisierter Gespräche, die sich alle leicht unterscheiden, aber in ihrer Aggregatwirkung gesellschaftliche Realitätsfragmentierung erzeugen. Traditionelle Monitoring-Ansätze, die nach Mustern koordinierter Kampagnen suchen, greifen nicht mehr.

¹⁵ Katja Muñoz, „The Seemingly Conscious AI Problem,“ Brave New Digital Space (Substack), 2. September 2025, <https://bravenewdigitalspace.substack.com/p/the-seemingly-conscious-ai-problem> (zuletzt abgerufen am 03.02.2026).

FRÜHWARNINDIKATOREN

- Langfristig aktive Accounts mit konsistentem Verhalten ohne klare menschliche Betreiber
- Open-Source Agent Frameworks Monitoring um Entwicklung von Agent-Systemen mit versteckten oder manipulierbaren Zielfunktionen in Developer-Communities zu erfassen
- Mapping und Monitoring des Companion-App-Markt
- Zunehmende Verwendung emotionaler/ persönlicher Sprache in KI-Systemen

QUICK WINS

- Kennzeichnungspflicht für KI-Agenten in öffentlichen digitalen Räumen
- Transparenzanforderungen für KI-Assistenten
- Monitoring-Programm für autonome Systeme in sozialen Netzwerken
- Registrierungspflicht für autonome Systeme mit Verhaltenszielen
- Behavioral Audits für KI-Assistenzsysteme (insbesondere Gesundheit, Finanzen, Politik)
- Entwicklung von Detection-Methoden für langfristige autonome Operationen (Forschungsförderung)
- Anti-Anthropomorphisierungs-Standards für hochsensitive Bereiche (Gesundheit, Finanzen)

DER WEG NACH VORN – STRATEGIEN FÜR DEUTSCHLAND

Das Problem der Analyse ohne Konsequenz

Forschungsinstitutionen, Sicherheitsbehörden und Tech-Unternehmen dokumentieren Bot-Netzwerke, kartieren Desinformationskampagnen und visualisieren koordinierte nicht authentische Kampagnen. Diese Arbeit führt unweigerlich auch zu einer unbequemen Erkenntnis: Die analytische Phase ist weitgehend abgeschlossen. Am Beispiel Russlands wird sichtbar, dass Taktiken breit dokumentiert, Techniken katalogisiert und Größenordnung bekannt sind. Weitere Analysen bringen kaum neue Erkenntnisse und produzieren oft Variationen bereits bekannter Muster.

Das eigentliche Problem liegt in strukturellen Blockaden, die verhindern, dass aus Analyse Handlung wird. Trotz existierender Plattformregulation sowie der EU-KI-Verordnung zeigen sich Tech-Unternehmen und Social-Media-Betreiber zunehmend unwillig, auf Basis externer Untersuchungen zu handeln.¹⁶ Sanktionen aufgrund von EU-Versstößen werden von der derzeitigen US-Regierung als Angriff interpretiert. Selbst wenn Plattformen handeln, werden gelöschte Bot-Accounts binnen Stunden wiederhergestellt. Ohne effektive Kooperation entwickelt sich das Ganze zu einem industrialisierten Whack-a-Mole-Spiel, das strukturell nicht zu gewinnen ist.

Präventionsmaßnahmen skalieren nicht. Pre-Bunking funktioniert in kontrollierten Studien,¹⁷ versagt jedoch bei der Geschwindigkeit feindlicher Operationen. In Medienkompetenz wird nicht ausreichend investiert. Ausnahmen bilden hier die baltischen Staaten, Finnland und Taiwan.¹⁸

Abschreckung existiert nicht und Attribution hat ihre Schärfe verloren. Russland operiert täglich im westlichen Informationsraum, während der umgekehrte Fall nicht existiert. Die demokratische Welt ist durch Ethikdebatten und Eskalationsängste gelähmt, während der Gegner keinerlei vergleichbaren Druck erfährt. Der Zeitpunkt zum Handeln ist jetzt. Abschreckung erfordert glaubhafte Drohung. Deutschland und Europa verabschieden Regulierungen, die sie nicht konsequent durchsetzt – Eskalationsangst und Verdrängung ersetzen strategisches Handeln. Auf dieser Basis sind

folgende Veränderungen notwendig, um diese Blockade zu lösen und Deutschland handlungsfähig zu machen.

Handlungsempfehlungen für Deutschland angesichts hybrider Gefahren

Strategische Neuausrichtung Richtung technologischer Souveränität ist Sicherheitspolitik

Deutschland muss technologische Souveränität als fundamentale Sicherheitspolitik begreifen, nicht lediglich als Industrieförderung. Wenn kritische Infrastruktur auf ausländisch kontrollierten Plattformen läuft, werden Maßnahmen gegen hybride KI-Bedrohungen systemisch unmöglich. Dies erfordert eine institutionelle Verankerung auf höchster Ebene.

Der Bundeskanzler muss technologische Souveränität zur Chefsache machen. Der Nationale Sicherheitsrat sollte KI-Infrastruktur als ständigen Tagesordnungspunkt etablieren und wie kritische Infrastruktur behandeln. Eine interministerielle Arbeitsgruppe zu „KI & Hybride Bedrohungen“ unter Federführung des Kanzleramtes sollte dringend ressortübergreifende Koordination sicherstellen.

Sofortmaßnahmen und Quick Wins umsetzen

Teil 2 identifiziert konkrete Sofortmaßnahmen: Transparenzpflichten, Infrastruktur-Mapping, Diversifizierungsanforderungen, Kennzeichnungspflicht für Agenten, Behavioral Audits. Diese Quick Wins sind binnen 6 bis 12 Monaten umsetzbar und müssen umgehend in Regierungshandeln übersetzt werden.

Die Bundesregierung muss verpflichtende Souveränitätsprüfungen für alle Digitalisierungsprojekte einführen. Kritische Systeme erfordern Nachweis begrenzter Abhängigkeiten, vorhandene Exit-Strategien und gewährleistetere Datensouveränität. Föderale digitale Fragmentierung sollte durch bundesweite Mindeststandards überwunden werden.

Investitionen und strukturelle Reformen zügig angehen

Das Zeitfenster für präventives Handeln schließt sich. Jeder Monat Verzögerung macht Infrastruktur-Lock-In irreversibler. Deutschland muss massiv in europäische KI-Infrastruktur-Alternativen investieren.

16 Siehe Beispiel Liz Alderman und Adam Satariano, „E.U. Hits Elon Musk’s X with \$140 Million Fine,” The New York Times, 5. Dezember 2025, <https://www.nytimes.com/2025/12/05/technology/eu-elon-musk-x-140-million-fine.html> (zuletzt abgerufen am 03.02.2026).

17 Annette Kroeber-Riel, „Prebunking: Desinformation bekämpfen, bevor sie verbreitet wird,” Google Blog, 25. April 2024, <https://blog.google/intl/de-de/unternehmen/inside-google/prebunking-desinformation-bekaempfen/> (zuletzt abgerufen am 03.02.2026).

18 James Brooks, „Finnish children learn media literacy at 3 years old. It’s protection against Russian propaganda,” Associated Press, 5. Januar 2026, <https://apnews.com/article/fake-news-classrooms-finland-russia-194b32d8829838bfe47469d6ff357689> (zuletzt abgerufen am 15.01.2026).

Medienkompetenz muss nach baltischem Vorbild massiv ausgebaut werden. Estland, Finnland und Taiwan zeigen, dass demokratische Gesellschaften resilient werden, wenn politischer Wille vorhanden ist.

Europäische Dimension erkennen und nutzen

Europäische Infrastruktur-Souveränität erfordert koordinierte Anstrengungen. Deutschland kann und sollte hier Treiber sein. Die Bundesregierung sollte eine EU-Initiative für digitale Souveränität anstoßen und konkrete Governance-Strukturen für autonome KI-Agenten etablieren. Dies erfordert nicht zwingend Einstimmigkeit aller 27 Mitgliedstaaten. Das Konzept des „pragmatic federalism“, bei dem willige EU-Mitglieder in so vielen Bereichen wie möglich gemeinsam voranschreiten, wird bereits erprobt.¹⁹ Deutschland sollte mit gleichgesinnten Partnern vorangehen und Standards setzen, denen andere folgen können. Warten auf europäischen Konsens bedeutet de facto Handlungsunfähigkeit.

Die Bedrohungsanalyse ist vorhanden, die Szenarien durchgespielt, die Quick Wins identifiziert. Was nun fehlt, ist der politische Wille, zügig den genannten Gefahren entgegenzutreten. **Nostalgie gegenüber dem vorherigen Status quo ist keine Strategie**, doch genau diese Sehnsucht scheint die Handlungsfähigkeit hierzulande zu lähmen. Deutschland und Europa müssen erkennen: Die regelbasierte Ordnung, in der hybride Bedrohungen Ausnahmen waren, existiert nicht mehr.

¹⁹ Martin Sandbu, „Europe Is Not as Weak as It Acts,“ Financial Times, 8. Februar 2026, <https://www.ft.com/content/52d2b108-7a9e-4558-abb3-920adc60bf24> (zuletzt abgerufen am 21.02.2026).

DGAP

Advancing foreign policy. Since 1955.

Rauchstraße 17/18
10787 Berlin
Tel. +49 30 254231-0
info@dgap.org
www.dgap.org
X @dgapev

Die Deutsche Gesellschaft für Auswärtige Politik e.V. (DGAP) forscht und berät zu aktuellen Themen der deutschen und europäischen Außenpolitik. Dieser Text spiegelt die Meinung der Autorinnen und Autoren wider, nicht die der DGAP.

Die DGAP ist gefördert vom Auswärtigen Amt aufgrund eines Beschlusses des Deutschen Bundestages.

Herausgeber
Deutsche Gesellschaft für
Auswärtige Politik e.V.

ISSN 2198-5936

Redaktion Jana Idris

Layout Luise Rombach

Design Konzept WeDo

Fotos Autorinnen und Autoren © DGAP



Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung – Nicht kommerziell – Keine Bearbeitungen 4.0 International Lizenz.