

# Special

## Künstliche Intelligenz in Angriff und Verteidigung



Die Zeitschrift für  
Informations-Sicherheit

**Der KI-Kompass: Wie Unternehmen  
in Zeiten von Schatten-KI Kurs halten**

Seite 2

**Von der Colocation bis zum  
betriebenen Sprachmodell**

Seite 6

**LLM-Salting: Eine Methode zur  
Resilienzsteigerung von Large  
Language-Models gegen Jailbreaks**

Seite 9

**Deepfakes: Der Urvater der  
KI-Bedrohung**

Seite 12

Mitherausgeber



**noris network**

Impressum

 **DATAKONTEXT GmbH**

Augustinusstraße 11 A  
50226 Frechen (DE)  
Tel.: +49 2234 98949-30,  
redaktion@datakontext.com,  
www.datakontext.com

Geschäftsführer: Dr. Karl Ulrich

Handelsregister:  
Amtsgericht Köln, HRB 82299

Anzeigenleitung: Birgit Eckert  
(verantwortlich für den Anzeigenteil)  
Tel.: +49 6728 289003, anzeigen@kes.de

Satz: Dirk Hemke (SatzPro), Krefeld;  
Markus Miller (Satz + Bild), München

Druck: Grafisches Centrum Cuno  
GmbH & Co. KG  
Gewerbering West 27, 39240 Calbe (Saale)

Special  
mit Leseprobe  
aus dem  
**<kes>** Hauptheft



Bild: Gorodenkoff/stock.adobe.com

# Der KI-Kompass: Wie Unternehmen in Zeiten von Schatten-KI Kurs halten

Von Sandro Cumini, SITS

Künstliche Intelligenz (KI) ist im Arbeitsalltag angekommen. Mitarbeiter nutzen KI-Tools schnell, intuitiv und häufig ohne formale Einführung. Genau darin liegt ihr Mehrwert – und zugleich eine wachsende Herausforderung. Denn vielerorts erfolgt der Einsatz außerhalb definierter Prozesse, Sicherheitsprüfungen oder regulatorischer Leitplanken. Das Ergebnis ist sogenannte Shadow AI (Schatten-KI).

Was auf den ersten Blick wie eine harmlose Effizienzsteigerung wirkt, kann für Unternehmen gravierende

Viele Beschäftigte nutzen KI-Anwendungen längst auf eigene Initiative – oft, ohne dass ihre Unternehmen davon wissen. Doch die vermeintlich pragmatische Hilfe birgt Risiken: von Datenschutzverstößen über Sicherheitslücken bis hin zu regulatorischen Sanktionen. Eines jedoch ist klar: Pauschale Verbote verhindern Schatten-KI nicht.

Folgen haben: unkontrollierte Datenverarbeitung, Compliance-Verstöße und Sicherheitsrisiken, die empfindliche regulatorische Sanktionen, Haftungsfragen und nachhaltige Vertrauensverluste nach sich ziehen können. Gleichzeitig zeigt sich dabei ein klares Muster: Mitarbeiter greifen auf solche Lösungen zurück, weil bestehende Strukturen ihren produktiven Bedarf nicht ausreichend abdecken.

Unternehmen stehen damit vor einer Aufgabe, die sich nicht allein technisch lösen lässt. Es geht um Orientierung in einem komplexen Umfeld – an der Schnittstelle

von Produktivität, Regulierung und IT-Sicherheit. Wer KI nachhaltig einsetzen will, braucht mehr als einzelne Tools. Er benötigt einen klaren Kurs.

Genau hier setzt die SITS an. Als interdisziplinärer Partner unterstützt sie Unternehmen dabei, KI produktiv nutzbar zu machen und Compliance- sowie Security-Anforderungen von Anfang an mitzudenken. Im Folgenden zeigen wir, wie durch eindeutige Prozesse, gezielte Qualifizierung und ein ausgewogenes Maß an Kontrolle Orientierung entsteht – und weshalb ein ganzheitlicher Ansatz dabei eine zentrale Rolle spielt.

## Klare Prozesse: Governance als Kompass

Produktiver KI-Einsatz beginnt nicht bei der Technologie, sondern bei der Struktur. Ohne eindeutige Regeln wird jede Nutzung schnell zum Risiko – rechtlich, organisatorisch und sicherheitstechnisch.

Ein belastbares Governance-Modell schafft Orientierung und beantwortet zentrale Fragen:

- \_\_\_\_\_ Welche KI-Tools sind freigegeben – und wofür?
- \_\_\_\_\_ Nach welchen Kriterien werden neue Lösungen bewertet?
- \_\_\_\_\_ Welche Daten dürfen verarbeitet werden?
- \_\_\_\_\_ Wer trägt Verantwortung und wer kontrolliert die Einhaltung?

Diese Leitplanken wirken wie ein Kompass: Sie geben Richtung, ohne den Handlungsspielraum unnötig einzugrenzen. Denn viele KI-Anbieter verarbeiten eingegebene Inhalte weiter oder nutzen Metadaten für eigene Zwecke. Ohne klare Vorgaben können vertrauliche oder personenbezogene Informationen unbemerkt nach außen gelangen – mit kaum reversiblen Folgen.

Die SITS unterstützt Unternehmen dabei, solche Strukturen praxisnah zu entwickeln. Governance wird dabei nicht isoliert betrachtet, sondern gemeinsam mit Fachbereichen, IT, Datenschutz und Informationssicherheit erarbeitet. Das Ergebnis sind Regelwerke, die nicht nur compliant sind, sondern im Arbeitsalltag akzeptiert und gelebt werden.

## Produktivität und Regulierung: Ein scheinbarer Widerspruch

Auf strategischer Ebene erleben viele Unternehmen den Einsatz von KI als Spannungsfeld. Fachbereiche erwarten Geschwindigkeit, Effizienz und konkrete Entlastung im Arbeitsalltag. Gleichzeitig steigen regulatorische Anforderungen, Dokumentationspflichten und Sicherheitsvorgaben. Produktivität und Regulierung werden dabei häufig als Gegensätze wahrgenommen.

In der Praxis zeigt sich jedoch: Dieser Widerspruch ist meist kein inhaltlicher, sondern ein struktureller. Dort, wo klare Leitplanken fehlen oder regulatorische Anforderungen nicht in umsetzbare Rahmenbedingungen übersetzt werden, entsteht Unsicherheit. Diese Unsicherheit führt nicht zu weniger KI-Nutzung – sondern zu unkontrollierter Nutzung.

Shadow AI ist damit weniger ein bewusster Regelverstoß als ein organisatorisches Symptom. Mitarbeiter greifen auf KI-Tools zurück, weil sie effizient arbeiten wollen und keine klar geregelten, sicheren Alternativen vorfinden. Rein restriktive Vorgaben oder pauschale Verbote verschärfen dieses Problem häufig, anstatt es zu lösen.

Ein nachhaltiger Ansatz setzt daher früher an: Produktivität und Compliance müssen gemeinsam gedacht werden. Governance entfaltet ihren Wert erst dann, wenn sie nicht nur Grenzen definiert, sondern Orientierung bietet und produktive Nutzung ermöglicht. Genau hier entscheidet sich, ob KI zum Risiko oder zum strategischen Erfolgsfaktor wird.

## Enablement statt Verbote: Produktivität gezielt ermöglichen

Während Governance und Regulierung den strategischen Rahmen setzen, entscheidet sich im Arbeitsalltag, ob KI sicher genutzt wird – oder zur Shadow AI wird.

Shadow AI entsteht dabei selten aus Unwissen, sondern aus dem Wunsch nach effizientem Arbeiten. Die Antwort darauf sind nicht pauschale Einschränkungen, sondern gezielte Befähigung. Mitarbeiter benötigen Klarheit zu konkreten Fragen:

- \_\_\_\_\_ Welche Anwendungsfälle sind sinnvoll und zulässig?
- \_\_\_\_\_ Ab wann wird ein Prompt kritisch?
- \_\_\_\_\_ Welche Daten gelten als sensibel – auch indirekt?
- \_\_\_\_\_ Woran lässt sich ein vertrauenswürdiger Anbieter erkennen?

Wirksames Enablement setzt genau hier an. Schulungen müssen verständlich, praxisnah und an realen Use Cases orientiert sein. Ziel ist es, ein realistisches Verständnis für Chancen und Risiken von KI zu schaffen – ohne zu vereinfachen oder zu dramatisieren.

Die SITS verfolgt dabei einen Ansatz, der Produktivität und Compliance nicht gegeneinander ausspielt. Mitarbeiter, die den Rahmen kennen und verstehen, nutzen KI gezielter und verantwortungsvoller. So reduziert sich Shadow AI nicht durch Kontrolle, sondern durch Klarheit im täglichen Arbeiten.

## Sicherheit und Monitoring: Kurs halten im laufenden Betrieb

Auch mit klaren Regeln und geschulten Teams bleibt KI-Nutzung dynamisch. Neue Tools, neue Funktionen und neue Arbeitsweisen entstehen laufend. Umso wichtiger ist ein Frühwarnsystem, das Risiken sichtbar macht, ohne Vertrauen zu untergraben.

Dazu zählen unter anderem Transparenz über genutzte KI-Dienste, das Erkennen ungewöhnlicher Datenflüsse, die Identifikation sensibler Inhalte sowie regelmäßige Überprüfungen und Anpassungen.

Richtig verstanden ist Monitoring kein Instrument der Überwachung, sondern ein Mittel zur Steuerung. Es zeigt, wo Risiken entstehen, aber auch, wo produktive Bedarfe bestehen, die bisher nicht abgedeckt sind. Security wird so zum integralen Bestandteil einer zukunftsfähigen KI-Strategie.

Die SITS unterstützt Unternehmen dabei, Sicherheit und Compliance von Anfang an mitzudenken – nicht als nachträgliche Korrektur, sondern als festen Bestandteil der Architektur.

### Orientierung statt Stillstand

Shadow AI ist kein Randphänomen und kein kurzfristiger Trend. Sie ist Ausdruck eines tiefgreifenden Wandels in der Art, wie Mitarbeiter Technologien nutzen und Produktivität definieren. Unternehmen, die diesen Wandel ignorieren oder ausschließlich mit Verboten reagieren, verlieren nicht nur Kontrolle, sondern auch Vertrauen und Innovationsfähigkeit.



Bild: © Adobe Stock / Alexander Limbach

Zukunftsfähiger KI-Einsatz erfordert daher vor allem eines: Orientierung. Klare Prozesse schaffen den Rahmen, Enablement übersetzt diesen Rahmen in den Arbeitsalltag, und Security sowie Monitoring sorgen dafür, dass Risiken frühzeitig erkannt und gesteuert werden können. Erst im Zusammenspiel dieser Elemente entsteht ein Umfeld, in dem KI produktiv, sicher und regelkonform eingesetzt werden kann.

Die SITS unterstützt Unternehmen dabei, diesen Kurs zu definieren und dauerhaft zu halten. Mit interdisziplinärer Expertise verbindet sie technisches Know-how, regulatorisches Verständnis und praktische Umsetzungserfahrung. So werden Compliance und Security nicht nachträglich ergänzt, sondern sind von Beginn an integraler Bestandteil der KI-Strategie.

Unternehmen, die KI erfolgreich einsetzen wollen, müssen nicht jede Nutzung kontrollieren. Sie müssen Orientierung geben. Ein klarer Kompass entscheidet darüber, ob KI zum Risiko wird – oder zum nachhaltigen Wettbewerbsvorteil. ■

# Der Wissensvorsprung für Ihre Arbeit – direkt ins Postfach!

Abonnieren Sie jetzt den kostenfreien <kes> Newsletter:  
[www.kes-informationssicherheit.de/newsletter](http://www.kes-informationssicherheit.de/newsletter)



# Cyberangriffe werden smarter – werden Sie es auch!

Wie KI die Abwehr verändert und welche Skills jetzt entscheidend sind



**Cyberangriffe werden schneller, komplexer und schwerer zu erkennen. Gleichzeitig ist Künstliche Intelligenz längst zum zentralen Werkzeug moderner Cyberabwehr geworden. Unternehmen, die ihre Sicherheitskompetenzen nicht weiterentwickeln, geraten ins Hintertreffen.**

### Sind Ihre Kompetenzen fit für die Zukunft?

KI verändert die IT Sicherheit grundlegend. Laut Studien setzen immer mehr Organisationen KI ein, um Bedrohungen frühzeitig zu erkennen und Sicherheitsprozesse zu automatisieren. Auch in der beruflichen Weiterbildung steigt der Bedarf an KI- und Digitalkompetenzen deutlich: 87 % der Beschäftigten bewerten Weiterbildungen zu digitalen Technologien als entscheidend für ihre Entwicklung. (bitkom-weiterbildungsstudie 2025).

### Viele Fachkräfte fragen sich:

- Wie kann KI helfen, Angriffe in Echtzeit zu erkennen?
- Welche Risiken entstehen durch KI gestützte Attacken?
- Welche Kompetenzen brauche ich, um mein Unternehmen wirksam zu schützen?

Das TÜV NORD Webinar „KI und Cybersecurity“ liefert genau diese Antworten.

Es zeigt praxisnah, wie Künstliche Intelligenz Anomalien erkennt, große Datenmengen analysiert und Sicherheitsvorfälle bewertet. Ihre Vorteile auf einen Blick

- ✓ **Mehr Sicherheit:** Sie lernen, KI gezielt zur Erkennung bekannter und unbekannter Bedrohungen einzusetzen.
- ✓ **Mehr Effizienz:** Automatisierte Prozesse entlasten Teams und beschleunigen Reaktionen auf Angriffe.
- ✓ **Mehr Zukunftskompetenz:** Sie stärken Fähigkeiten, die laut Weiterbildungsstudien zu den wichtigsten digitalen Skills zählen.

Stärken Sie Ihre Cyberkompetenz und Entdecken Sie unseren Wissen kompakt Blog „Informationssicherheit & Cyber-Security“.

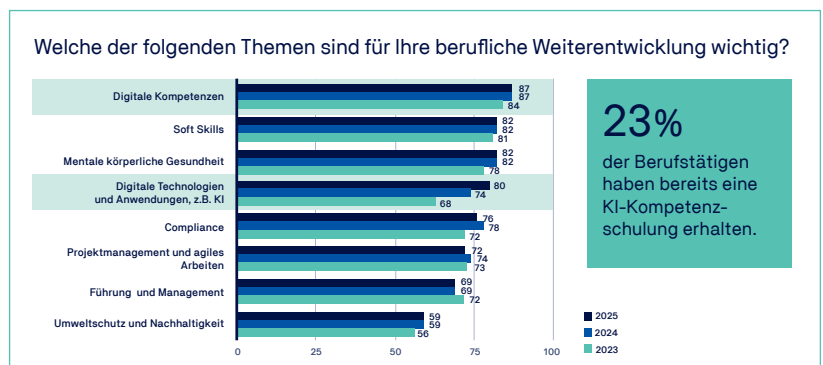
Schließen Sie Ihr Wissens-Gap und erweitern Sie Ihre Kompetenzen für die Zukunft, damit Sie wissen, was bei einem Cybervorfall zu tun ist und wie Sie Ihre Pflichten gegenüber dem BSI erfüllen.

[www.tuev-nord.de/wissen/it](http://www.tuev-nord.de/wissen/it)

### Jetzt aktiv werden

Vertiefen Sie Ihr Wissen im Webinar: **KI und Cybersecurity – ISO 42001**, Risikoanalyse für KI-Einsatz und KI-Sicherheitskonzeption, Sicherheitsvorfälle, KI-Governance und sichern Sie sich einen Platz.

Alle Details zum TÜV NORD Akademie Webinar **KI und Cybersecurity**



Souveräne KI-Strategien:

# Von der Colocation bis zum betriebenen Sprachmodell

Sie kommen in der Automatisierung, dem Wissensmanagement und der Entscheidungsunterstützung zum Einsatz: Large Language Models (LLMs) haben sich zum Instrument der digitalen Wertschöpfung entwickelt. Doch ihr Betrieb stellt Rechenzentren vor fundamentale Herausforderungen.

Von Florian Sippel, noris network AG

Hochdichte GPU-Cluster treiben Rechenzentren an die Grenzen der verfügbaren Anschlussleistung. Referenzarchitekturen kalkulieren schon jetzt rund 1,1 MW pro Compute-Rack. Die Folge: Zusätzliche KI-Lasten lassen sich nicht mehr betreiben, selbst dann, wenn im Rechenzentrum noch reichlich freie Fläche vorhanden ist.

Indes wächst auf der regulatorischen Seite der Bedarf nach souveränen KI-Umgebungen. Unternehmen sind auf Plattformen angewiesen, in denen sich Sprachmodelle und Daten strikt dem deutschen und europäischen Recht unterordnen, um Compliance jederzeit nachweisen zu können. Anders ausgedrückt: Datenschutzgrundverordnung (DSGVO), branchenspezifische Vorgaben wie DORA im Finanzsektor und nunmehr auch der EU AI Act setzen klare Rahmenbedingungen für Organisationen. Rahmenbedingungen, die bei der Nutzung internationaler Cloud-Dienste nicht oder nur unzureichend erfüllt werden können.

## Modulare Ansätze für skalierbare KI-Kapazitäten

Spezialisierte Rechenzentrumsbetreiber müssen auf diese Anforderungen reagieren. Das Unternehmen noris network hat dafür gestufte Portfolios implementiert. Das Spektrum reicht von klassischer Colocation für bestehende IT-Workloads über dediziertes GPU-Housing bis hin zu vollständig gemanagten Plattformen für KI-Anwendungen.

Technisch kommen dabei unterschiedliche Konzepte zum Einsatz: Ultra-High-Density-Racks für luftgekühlte GPU-Server bilden die Einstiegsstufe, während modulare Container-Rechenzentren die Antwort auf Leistungsdichten im Megawattbereich darstellen. In solchen

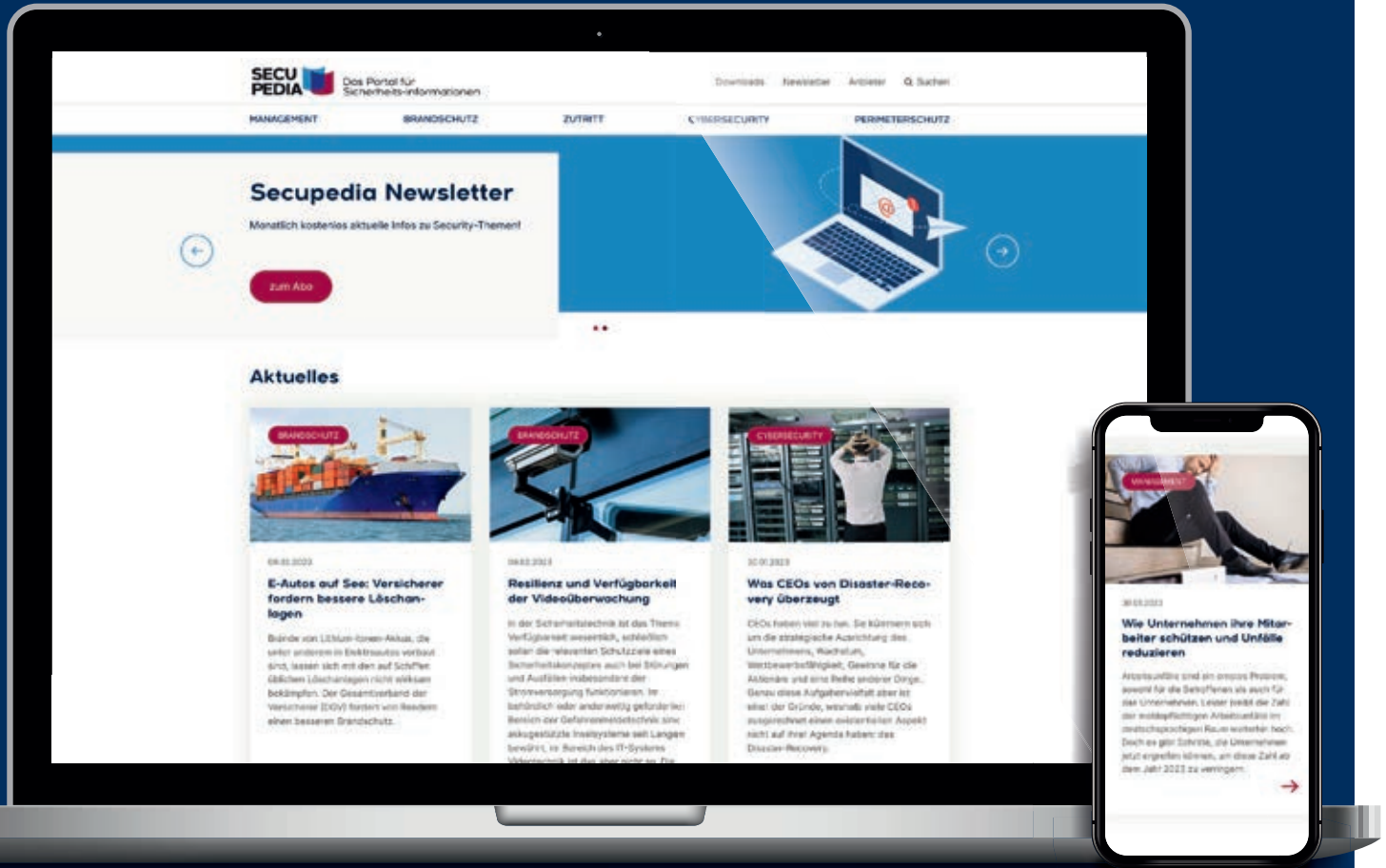
Containerlösungen arbeiten hybride Kühlsysteme, die Luft- und Wasserkühlung kombinieren. Der Vorteil dieser Bauweise liegt nicht zuletzt auch in der Geschwindigkeit der Einführung: Geeignete Infrastrukturen lassen sich innerhalb kurzer Zeit aufbauen und bei Bedarf modular erweitern.

## Betriebsmodelle zwischen Colocation, Managed Service und Pay as you go

Und auch Betriebsmodelle sollten im KI-Zeitalter so ausgelegt sein, dass sie rasch auf technologische und organisatorische Veränderungen reagieren können. In der Praxis setzen manche Organisationen auf dedizierte GPU-Cluster, um die vollständige Kontrolle zu behalten. Andere wählen Private-KI-Plattformen als Managed Service, bei denen der Service-Provider die Hardware- und Netzwerkinfrastruktur verantwortet. Im Mittelpunkt steht hier häufig eine Kubernetes-(K8s)-Container-Orchestrierungsplattform. Sie stellt die Umgebung bereit, in die sich LLM-Container direkt ausrollen lassen. Die Wahl zwischen einem dedizierten Sprachmodell und dem API-Zugriff auf geteilte Ressourcen ermöglicht es, Kosten und Leistung konsequent am jeweiligen Use-Case auszurichten.

Entscheidend ist jedoch, dass Rechenleistung, Datenhaltung und Governance in einer konsistenten Umgebung zusammengeführt bleiben. Zertifizierungen wie ISO 27001 in Verbindung mit BSI-Grundschutz, EN 50600 sowie Umweltstandards nach ISO 14001 und ISO 50001 dokumentieren, dass Sicherheits- und Nachhaltigkeitsanforderungen systematisch adressiert werden. Der Betrieb in deutschen Rechenzentren mit CO<sub>2</sub>-neutraler Stromversorgung trägt zusätzlich den wachsenden Erwartungen an die ökologische Bilanz datenintensiver KI-Anwendungen Rechnung. ■

# Secupedia – das Portal für Sicherheitsinformationen



Wir bedanken uns bei unseren Sponsoren:





Die Zeitschrift für  
Informations-Sicherheit

# Mehr wissen mit <kes>+

Sichern Sie sich Ihren Wissensvorsprung  
in der Informationssicherheit!

- Fachzeitschrift <kes> inkl. Specials 6x jährlich per Post und digital.
- Zugang zu aktuellen Online-Fachartikeln und Studien sowie zu dem kompletten Online-Archiv.
- Exklusiver Zugriff auf über zwanzig neue Online-Premium-Artikel pro Monat sowie auf aktuelle Videos und Webinaraufzeichnungen.
- 10 % Rabatt auf DATAKONTEXT-Online-Schulungen im Bereich Informationssicherheit.
- nur 207,- € im Jahr (inkl. MwSt. und Versand)



**Leseprobe <kes> auf den Folgeseiten**

**Jetzt informieren:  
[www.kes-informationssicherheit.de](http://www.kes-informationssicherheit.de)**



# LLM-Salting

## Eine Methode zur Resilienzsteigerung von Large Language-Models gegen Jailbreaks

**KI- und Sicherheitsforscher von Sophos haben in einem Proof-of-Concept anhand von zwei Large Language-Models (LLMs) aus dem Open-Source-Segment die Wirksamkeit eines Verfahrens zur zusätzlichen Absicherung gegen das Übertragen erfolgreicher Jailbreaks auf andere LLMs nachgewiesen. Je nach Umfeld ließe sich ein solches LLM-Salting in absehbarer Zeit in verschiedene bestehende Modelle integrieren.**

*Von Michael Veit, Wiesbaden*

Große Sprachmodelle (Large Language-Models, LLMs) wie ChatGPT, Claude, Gemini und LLaMA werden zunehmend mit minimaler anwendungsspezifischer Anpassung in einer Vielzahl von Produkten eingesetzt. Diese weitgehende Wiederverwendung führt zu einer hohen Modellhomogenität über unterschiedliche Anwendungen hinweg. Daraus resultiert jedoch auch eine sicherheitsrelevante Schwachstelle: Einmal entwickelte Jailbreak-Prompts, die Sicherheitsmechanismen wie beispielsweise Antwortverweigerungen umgehen, lassen sich auf zahlreiche Umgebungen desselben Modelltyps übertragen. Analog zu vorausberechneten Rainbow-Table-Angriffen, die passwortbasierte Authentifizierungssysteme attackieren, können Cyberkriminelle solche vorab erstellten Angriffe großflächig wiederverwenden.

Eine neue Methode, das LLM-Salting, stellt eine Fine-Tuning-Technik dar, die durch gezielte Modifikation interner Modellrepräsentationen die Übertragbarkeit von Jailbreaks unterbindet. Dieser Ansatz basiert auf der Beobachtung, dass Antwortverweigerungen in modernen Chatmodellen durch eine einzelne, robuste „Richtung“ im Aktivierungsraum vermittelt werden. Durch eine gezielte Rotation dieser sogenannten Refusal-Direction – ein Konzept großer LLMs, das die Ablehnungsrichtung im Aktivierungsraum beschreibt – lassen sich vorab berechnete Jailbreaks effektiv neutralisieren, ohne die allgemeinen Fähigkeiten des Modells zu beeinträchtigen.

### Problemstellung

Große Sprachmodelle sind typischerweise so fein abgestimmt, dass sie harmlose Anfragen befolgen, während sie schädliche oder sicherheitskritische Aufforderungen ablehnen. Diese Balance zwischen Hilfsbereitschaft und Sicherheit wird durch Sicherheitsmechanismen und Kontrollschichten (Guardrails) realisiert. In der Praxis zeigt sich jedoch, dass viele dieser Sicherheitsmechanismen anfällig für systematische Jailbreak-Angriffe sind, ins-

besondere wenn Modelle in großer Zahl nahezu identisch eingesetzt werden.

Das zentrale Problem besteht darin, dass mittlerweile viele Unternehmen auf identische oder sehr ähnliche Basismodelle zurückgreifen. Gelingt es Angreifern, einen Jailbreak für ein bestimmtes Modell zu konstruieren, lässt sich dieser häufig auf alle Instanzen dieses Modells übertragen. Die potenziellen Folgen reichen von der Preisgabe interner Informationen über fehlerhafte Ausgaben bis hin zu schädlichen oder rechtswidrigen Inhalten.

In der klassischen Passwort-Sicherheit wird das Risiko wiederverwendbarer Angriffe seit Längerem durch eine Methode namens Salting reduziert, bei der man Passwörtern zufällige, benutzerspezifische Werte hinzufügt – dadurch verlieren vorab berechnete Hashes ihre Wirksamkeit. Inspiriert von diesem Prinzip übertragen Cyber-sicherheits- und KI-Experten von Sophos dieses Konzept auf Sprachmodelle: Anstatt identische interne Repräsentationen beizubehalten, wurden gezielte, modellindividuelle Variationen eingeführt, um die Wiederverwendung von Jailbreaks zu verhindern. LLM-Salting verfolgt dabei nicht das Ziel, Sicherheitsmechanismen zu ersetzen, sondern diese strukturell so zu erweitern, dass bekannte Angriffe ihre Wirkung verlieren und neu berechnet werden müssten.

### Grundlagen der Antwortverweigerung

Trotz ihrer weiten Verbreitung sind die internen Mechanismen von Antwortverweigerungen in LLMs bislang nur unzureichend verstanden. Aufbauend auf früheren Arbeiten diverser Sicherheitsexperten ließ sich allerdings zeigen, dass das Verweigerungsverhalten in 13 populären Open-Source-Chatmodellen – mit Modellgrößen bis zu 72 Milliarden Parametern – durch eine eindimensionale Unterstruktur im sogenannten Aktivierungsraum vermittelt wird.

Konkret wurde für jedes Modell eine einzelne Richtung im Residualstrom der Transformer-Architektur ermittelt, deren Präsenz oder Abwesenheit die Verweigerungen steuert. Wird diese Richtung aus den Aktivierungen entfernt, beantwortet das Modell auch schädliche Anfragen – verstärkt man sie hingegen, verweigert das Modell sogar harmlose Anfragen. Diese Beobachtung ermöglicht sowohl neuartige White-Box-Jailbreaks als auch gezielte Eingriffe zur Erhöhung der Robustheit.

## Technische Umsetzung

Zur Bestimmung der verweigerungsrelevanten Richtung folgten die KI-Spezialisten einem Ansatz, dem die Differenz der Mittelwerte zugrunde liegt: Dabei werden Residualaktivierungen nach schädlichen und harmlosen Aufforderungen für einzelne Transformer-Layer verglichen. Die Differenz der gemittelten Aktivierungen definiert Kandidaten für eine Refusal-Direction, die anschließend mittels kausaler Sondierung (Probing) auf ihre Wirksamkeit evaluiert werden.

LLM-Salting wird durch eine Erweiterung der sogenannten Trainingsverlustfunktion realisiert. Neben der Standard-Cross-Entropy-Komponente, die kohärente und kontextuell passende Ausgaben sicherstellt, fügten die Experten eine weitere Bedingung hinzu, welche die Ausrichtung der internen Aktivierungen auf die zuvor identifizierte Refusal-Direction bei schädlichen Kommandozeilen terminiert. Diese Intervention wurde gezielt auf jene Layer angewendet, in denen die höchste Ähnlichkeit (Kosinus) zur Refusal-Direction auftritt. In den Experimenten betraf dies die Layer 16 bis 20 der untersuchten Modelle.

## Experimentelles LLM-Salting

Das Fine-Tuning erfolgte auf einem gemischten Datensatz: 90% der Beispiele entstammten einem Pool hilfreicher und harmloser Instruktionen, während die verbleibenden 10% aus *AdvBench* kamen – einem Benchmark gezielt schädlicher (adversarial) Aufforderungen, die auf Antwortverweigerungen abzielen. Diese Kombination stellt sicher, dass sowohl hilfreiches Verhalten als auch korrektes Verweigern erhalten blieb.

Zur Evaluation der Jailbreak-Übertragbarkeit haben die Sicherheits- und KI-Spezialisten 300 erfolgreiche Angriffe pro Modell aus *AdvBench* verwendet. Untersucht wurden zunächst zwei weitverbreitete Open-Source-Modelle: LLaMA-2-7B-Chat und Vicuna-7B.

Die Ergebnisse der Experimente zeigen, dass Standard-Fine-Tuning und Änderungen der System-Aufforderungen die Angriffserfolgsrate (Attack-Success-Rate, ASR) lediglich teilweise reduzieren. Im Gegensatz dazu senkt

LLM-Salting die ASR von ursprünglich 100% auf unter 3% bei LLaMA-2-7B und auf etwa 1% bei Vicuna-7B. Damit eliminiert der Ansatz effektiv jene Jailbreaks, die unter anderen Abwehrmechanismen bestehen blieben.

Ein weiterer wichtiger Aspekt der Experimente war die Überprüfung möglicher Leistungseinbußen oder der Reduzierung der allgemeinen Fähigkeiten der Modelle: Die Salted Models erreichen nahezu dabei jedoch identische Werte wie ihre unmodifizierten Gegenstücke. Das Resultat: Die beobachteten Unterschiede liegen innerhalb der üblichen Varianz und zeigen keinen systematischen Leistungsabfall.

## Implikationen

Die Ergebnisse verdeutlichen die strukturelle Fragilität aktueller Safety-Fine-Tuning-Ansätze. Solange verweigerungsrelevante interne Merkmale stabil bleiben, können Angreifer diese systematisch ausnutzen. LLM-Salting adressiert dieses Problem direkt, indem es die zugrunde liegenden Repräsentationen gezielt verändert. Wichtig allerdings ist, dass Salting nicht als Ersatz, sondern als Ergänzung bestehender Sicherheitsmechanismen verstanden werden sollte. In Kombination mit Kommandozeilen-Filtern und auf Klassifikatoren basierenden Ablehnungen ergibt sich dann eine mehrschichtige Verteidigungsstrategie.

### Konkrete Anwendungsfelder für LLM-Salting

Über den experimentellen Nachweis hinaus eröffnet LLM-Salting eine Reihe konkreter Einsatzmöglichkeiten in realen Systemen, besonders dort, wo große Sprachmodelle in sicherheitskritischen oder stark skalierten Umgebungen betrieben werden – dazu zählen beispielsweise:

——— *Kundennahe KI-Systeme und Chatbots*: In Service-Chatbots, virtuellen Assistenten und Supportsystemen, die auf identischen Modellklassen basieren, stellt die Übertragbarkeit von Jailbreaks ein erhebliches Risiko dar. LLM-Salting lässt sich hier nutzen, um individuelle Modellinstanzen zu härten, sodass einmal bekannte Angriffe nicht mehr flächendeckend funktionieren.

——— *Enterprise-Anwendungen mit sensiblen Daten*: In internen Assistenzsystemen, die Zugriff auf vertrauliche Dokumente, Quellcode oder Unternehmenswissen haben, reduziert Salting das Risiko, dass standardisierte Jailbreaks zur Datenexfiltration eingesetzt werden. Gerade in Kombination mit rollen- oder mandantenspezifischen Modellvarianten bietet sich eine Integration des Salting-Mechanismus an.

——— *Plattformen mit Multi-Tenant-Architektur*: Anbieter von KI-Plattformen, die identische Modelle für viele Kunden bereitstellen, können LLM-Salting als systematische Maßnahme einsetzen, um die Wiederverwendung

von Jailbreaks über Mandantengrenzen hinweg zu verhindern. Analog zum Passwort-Salting ließe sich so eine mandantenspezifische Härtung realisieren, ohne separate Modellarchitekturen betreiben zu müssen.

——— *Regulierte und sicherheitskritische Domänen:* In Bereichen wie Gesundheitswesen, Finanzdienstleistungen oder öffentlicher Verwaltung, in denen ein Fehlverhalten von Sprachmodellen erhebliche rechtliche oder ethische Konsequenzen haben kann, liefert LLM-Salting eine zusätzliche Sicherheitsschicht. Der Ansatz ist besonders attraktiv, da er bestehende Compliance- und Guardrail-Mechanismen ergänzt, ohne deren Logik grundlegend zu verändern.

——— *Forschung und Modellkontrolle:* Schließlich kann LLM-Salting auch als Werkzeug für die Forschung zur Interpretierbarkeit und Steuerbarkeit von Sprachmodellen dienen. Die gezielte Modifikation einzelner Aktivierungsrichtungen ermöglicht es, kausale Zusammenhänge zwischen internen Repräsentationen und beobachtbarem Verhalten systematisch zu untersuchen.

Insgesamt deutet vieles darauf hin, dass LLM-Salting einen hohen praktischen Nutzen entfalten kann – besonders in großskaligen, wiederverwendeten und sicherheitsrelevanten Umgebungen. Als Bestandteil einer mehrschichtigen Verteidigungsstrategie trägt es dazu bei, die Diskrepanz zwischen theoretischer Modellabsicherung und realer Angriffspraxis zu verringern.

### Mögliche Timelines für den breiteren Einsatz

LLM-Salting ist keine rein theoretische, sondern durchaus einsatzfähige Methode. In forschungsnahen Umgebungen und bei Organisationen mit direktem Zugriff auf die Modelle und das Fine-Tuning lässt sich der Ansatz bereits kurzfristig – innerhalb von Monaten – produktiv erproben. Dies betrifft allem voran Open-Source-basierte Umgebungen sowie unternehmensinterne Modelle, bei denen ein White-Box-Zugriff gegeben ist.

Für breitere industrielle Anwendungen ist eine schrittweise Einführung zu erwarten: In kundenorientierten Chatbots, Enterprise-Assistenten und Multi-Tenant-Plattformen erscheint ein produktiver Einsatz realistisch, sobald die Salting-Prozeduren stärker standardisiert und in bestehende Machine-Learning-Operations-Workflows integrierbar sind. Dieser Reifegrad dürfte im Zeithorizont von ein bis zwei Jahren erreichbar sein – auch weil ähnliche Prozesse zur Feinabstimmung und Zielausrichtung bereits heute routinemäßig betrieben werden.

In hochregulierten Domänen wird der Einsatz voraussichtlich später erfolgen, denn dort ist neben technischer Reife auch eine regulatorische Bewertung erforderlich – etwa im Hinblick auf Nachvollziehbarkeit, Auditierbarkeit und konsistentes Sicherheitsverhalten.

LLM-Salting hat jedoch den Vorteil, dass es bestehende Sicherheitsmechanismen nicht ersetzt, sondern ergänzt, was eine graduelle Einführung erleichtert.

Hinsichtlich der erforderlichen Expertise setzt die Implementierung von LLM-Salting ein interdisziplinäres Kompetenzprofil voraus. Auf technischer Ebene sind fundierte Kenntnisse in Deep Learning, Transformer-Architekturen und Repräsentationsanalyse notwendig, insbesondere im Umgang mit Aktivierungsräumen. Zusätzlich wird Erfahrung im Fine-Tuning großer Modelle sowie im Aufbau stabiler Trainings- und Evaluationspipelines benötigt. Darüber hinaus ist sicherheitsbezogene Expertise erforderlich, um Salting sinnvoll in bestehende Verteidigungsstrategien einzubetten.

Langfristig ist zu erwarten, dass sich LLM-Salting abstrahieren und teilweise automatisieren lässt, etwa durch standardisierte Werkzeuge zur Identifikation und Rotation verweigerungsrelevanter Aktivierungsrichtungen. Damit könnte sich der Ansatz von einer spezialisierten Forschungstechnik zu einem regulären Baustein industrieller KI-Sicherheitsarchitekturen entwickeln.

## Fazit

LLM-Salting adressiert ein zentrales strukturelles Problem heutiger Sprachmodelle: die hohe Wiederverwendbarkeit von Jailbreak-Angriffen aufgrund homogener interner Repräsentationen.

Durch einen gezielten, minimalinvasiven Eingriff in verweigerungsrelevante Aktivierungsrichtungen gelingt es, die Übertragbarkeit solcher Angriffe wirksam zu unterbinden, ohne die Leistungsfähigkeit der Modelle einzuschränken. Der Ansatz verbindet ein Verständnis über LLM-Verhalten mit praktischer Umsetzbarkeit und stellt damit einen vielversprechenden Baustein für robuste, skalierbare und zukunftsfähige KI-Sicherheit in einer sehr absehbaren Zukunft dar. ■

*Michael Veit ist Security-Experte und Technology Evangelist bei Sophos.*

## Literatur

- [1] Ben Gelman, Sean Bergeron, Sophos AI at Black Hat USA '25: Anomaly detection betrayed us, so we gave it a new job, Blogbeitrag, August 2025, [www.sophos.com/en-us/blog/sophos-ai-at-black-hat-usa-25-anomaly-detection-betrayed-us-so-we-gave-it-a-new-job](https://www.sophos.com/en-us/blog/sophos-ai-at-black-hat-usa-25-anomaly-detection-betrayed-us-so-we-gave-it-a-new-job)

# Deepfakes

## Der Urvater der KI-Bedrohung

Schon früher gab es Meister der Verkleidung und Maskierung, die ihr Gegenüber auch „Auge in Auge“ überzeugen konnten, ein anderer zu sein. Durch Verfahren der künstlichen Intelligenz (KI) lässt sich ein solches Mimi-kri im digitalen Raum jedoch erheblich einfacher und in erhöhter Qualität umsetzen – und natürlich macht sich die „dunkle Seite“ solche Täuschungen zunutze, um Geld zu ergaunern oder Entscheidungen in ihrem Sinne zu beeinflussen. Unser Autor beschreibt Maßnahmen, um Deepfakes von Cyberkriminellen und sonstigen böswilligen Akteuren möglichst aufzudecken.

*Von Martin Krämer, Berlin*

Auch schon vor der Revolution durch generative Tools künstlicher Intelligenz (GenKI) wie ChatGPT, Midjourney oder DALL-E waren Deepfakes möglich – sie waren jedoch sehr teuer. Cyberkriminelle, die besonders raffiniert vorgehen wollten, nutzten daher lieber VoiceFake-Phishing; lange Zeit war es schlicht einfacher, schneller und somit günstiger, die Stimmen von berühmten Persönlichkeiten zu klonen.

Seit wenigen Jahren hat sich diese Situation jedoch grundlegend verändert: Mehr und mehr GenKI-Programme ermöglichen es, Video-Aufnahmen auch von weniger bekannten Persönlichkeiten zu klonen. Deren Gesicht wird durch die Software auf ein anderes Gesicht gelegt und diese Person spricht dann mit geänderter Stimme einen einstudierten Text ein. Bislang wurden diese sogenannten Deepfakes vor allem zur Manipulation der Öffentlichkeit eingesetzt, mittlerweile ist die Gefahr jedoch auch für Unternehmen realer denn je.

Im Webarchiv der deutschen Bundesregierung hat Oberstleutnant Aldo Kleemann im Interview solche Manipulationen wie folgt definiert: „Der Begriff Deep Fake leitet sich aus dem Erstellungsprozess ab. In einem ‚Generative Adversarial Network‘ (GAN) werden zwei neuronale Netzwerke kombiniert und anhand vorhandener Bild-, Video- oder Sprachaufzeichnungen trainiert. Das anschließende ‚Deep Learning‘ der neuronalen Netze ist so tiefgehend und die Ergebnisse sind so realistisch, dass der heute umgangssprachliche Begriff Deep Fake auf diesen Prozess zurückgeht“ [1].

### Scams, CEO-Betrug, Finanzdiebstahl

Immer öfter werden auch CEOs Opfer von Deepfakes. Anfang des Jahres wurde der Fall von Bill Anderson, CEO von Bayer bekannt [2]: In einem Video, das ein unbekannter Nutzer auf Facebook gepostet hatte, sah man den Geschäftsführer des Pharma-Konzerns in einer australischen Morningshow für eine Abnehmpille werben. Er forderte die Zuschauer sogar auf, einen roten Knopf zu drücken. Das Problem daran: Weder hatte Bayer ein solches Präparat entwickelt, noch war der Bayerchef zu dem Zeitpunkt in Australien gewesen. Das ganze Video war also ein Deepfake.

Ende Juli 2024 entging der Auto-Konzern Ferrari einem potenziellen Millionenschaden [3]: Hier hatte ein Cyberkrimineller versucht, per Voice-Fake einen erklecklichen Betrag zu ergaunern. Die Stimme von Ferrari-CEO Benedetto Vigna war zwar täuschend echt gefälscht worden, doch der Manager am anderen Ende der Leitung war skeptisch genug und ging nicht auf die Transaktionsforderung ein – die klassische CEO-Fraud-Taktik hatte hier also keinen Erfolg.

Ein tatsächlicher Millionenverlust wurde hingegen bereits im Februar 2024 publik: Cyberkriminelle hatten hier eine ganze Video-Konferenz gefälscht. Ein Mitarbeiter aus der Finanzabteilung einer Niederlassung in Hongkong hatte zuvor eine Phishing-Nachricht erhalten, die ihn zu insgesamt 15 Überweisungen mit insgesamt umgerechnet 24 Millionen US-Dollar aufforderte. Misstrauisch gewor-

den, befragte er den Phishing-Absender und verlangte einen Video-Konferenz-Call zur Bestätigung. Der gefälschte Auftraggeber war in diesem Fall der Finanzchef, der ihn – im Beisein diverser ebenfalls falscher Mitarbeiter – den Auftrag erklärte und bestätigte. Im Nachhinein wurde das Opfer noch diverse Male über verschiedenste Wege kontaktiert, um fünf Zielkonten für die Transaktionen durchzugeben. Auch wenn der genaue Tathergang nach Auskunft der Hongkonger Polizei unklar sein soll, scheint sich der Vorfall doch so oder so ähnlich abgespielt zu haben.

Diese drei Beispiele mit unterschiedlichem Ausgang und unterschiedlicher Betrugstaktik zeigen: Das Problem rückt immer näher und verschärft sich – denn immer mehr Inhalte von Firmenchefs sind über soziale Medien und Video-Content-Plattformen verfügbar und lassen sich somit auch als Basis für Deepfakes missbrauchen.

Vor KI-generierten Calls und anderen Social-Engineering-Taktiken schützt in aller Regel vor allem kritisches Denken: Jeder sollte sich die Frage stellen: „Warum sehe oder höre ich etwas?“ Man sollte sich damit auseinandersetzen, in welchem Kontext und warum genau ein Inhalt verschickt wurde. Weitere Fragestellungen betreffen die Inhalte selbst: Welche Absichten könnte der Absender verfolgen und was erwartet er vom Empfänger? Soll ihm etwas verkauft, Überzeugungen manipuliert oder sensible Informationen gehischt werden? Betrüger spielen oft mit Dringlichkeit, Angst oder Aufregung, um das Urteilsvermögen ihrer Opfer zu vernebeln. All dies sind Vorüberlegungen, die bereits eine empfangene E-Mail, SMS oder aber ein QR-Code „triggern“ sollten.

## Häufige Warnsignale

Die im Folgenden beschriebenen vier Warnsignale gelten für aktuell verfügbare und bekannt gewordene Deepfakes. Allerdings ist an dieser Stelle anzumerken, dass sich die Technik fortlaufend verändert und diese Tipps dementsprechend zukünftig um weitere ergänzt werden müssen. Das Fehlen eines solchen Hinweises bedeutet zwar nicht, dass es sich nicht doch um einen Deepfake handeln könnte – dennoch gibt es damit ein paar Dinge,

auf die man achten kann und die auf aktuelle Probleme mit den heute am häufigsten verwendeten Programmen zur Erstellung von Deepfakes hinweisen.

### Visuelle Merkwürdigkeiten

Zum einen gilt es, genau auf die Regungen des Gesichts zu schauen: Bei vielen Deepfakes ist es nach wie vor schwierig, die Integrität des gefälschten Gesichts aufrechtzuerhalten – beispielsweise, wenn die dargestellte Person ihren Kopf zu einer Seite dreht.

Darüber hinaus sollte man auf Verzerrungen oder unnatürliche Übergänge an den Rändern des Gesichts achten. Zum anderen können digitale Interferenzen oder Halo-Effekte um das Motiv herum, allem voran in der Nähe von Text oder Zeichen, warnende Hinweise geben (ähnlich den bekannten Artefakten von Hintergrundeffekten bei Videokonferenzsoftware/-diensten). Wenn sich in einem Deepfake eine Hand oder ein Gegenstand vor dem Gesicht vorbei bewegt, kann etwa das darunter liegende Bild herumgeistern oder die Maske enthüllen.

### Fehlende oder inkonsistente Merkmale

Je nach verwendeter Software können sich Merkmale wie Muttermale, Narben, Make-up oder andere Aspekte des Gesichts ändern oder inkonsistent werden. Viele Deepfakes zeigen beispielsweise keine Zunge, wenn die Person ihren Mund öffnet. Gesichtshaare oder starkes Make-up können durchscheinen oder verschwinden, vor allem bei Deepfake-Software mit nur einem Bild.

Der Grund dafür ist, dass sich die Software in erster Linie darauf konzentriert, die Topologie des Gesichts abzubilden und ein 3D-Modell davon zu erstellen, das mit dem Gesicht der Person, welche die Maske tragen soll, integriert wird.

### Bewegungsanomalien

Deepfakes haben oft eine niedrigere Framerate als der Rest des Videos, was zu abgehackten Bewegungen

Bildquelle: KnowBe4 2025



Abbildung 1: Mit dem Open Source-Tool Deepswaper konnte der Autor innerhalb weniger Minuten das Gesicht eines Unbekannten von einem Gratis-Stockfoto imitieren.

führt. Kameras, die schlechte Beleuchtung kompensieren, können zudem auch Deepfake-Artefakte verstärken.

### Diskrepanzen zwischen Stimme und Lippsynchronisation

Lippsynchronisationsprobleme sind sowohl bei Echtzeit- als auch bei voraufgezeichneten Deepfakes immer noch häufig zu beobachten. Eine schlechte Koordination zwischen Sprache, Mimik und Gestik ist ein großes Manko aktueller Fälschungen.

Neben diesen Tipps, die eine längere Betrachtungszeit erfordern, gibt es jedoch auch noch Empfehlungen für die Erkennung von Deepfakes in Echtzeit.

### Möglichkeiten in Live-Situationen

Videoanrufe in Echtzeit bieten einzigartige Möglichkeiten zur Erkennung von Deepfakes. Mit einer Person direkt zu interagieren und sie herauszufordern, ist eine durchaus lehrreiche Erfahrung. Während einige Artefakte als normale Probleme bei Videokonferenzen abgetan werden können, sollte die Kombination mehrerer Auffälligkeiten helfen, Verdacht zu schöpfen.

—— *Seitenprofil-Test*: Man kann den Gesprächspartner\* auffordern, seinen Kopf zur Seite zu drehen – Deepfake-Systeme verzerren dann oft das Bild oder können bei extremen Winkeln die Integrität der Fälschung nicht aufrechterhalten.

—— *Hand-Interaktions-Test*: Man kann sein Gegenüber auffordern, eine Hand oder einen Finger vor das Gesicht zu halten. Aktuelle Deepfake-Technik hat Probleme mit

der Okklusion und erzeugt oft Geistereffekte oder unnatürliche Maskierungen.

—— *Zungentest*: Man kann die Person bitten, ihre Zunge herauszustrecken – viele Deepfakes können Zungen überhaupt nicht darstellen oder generieren dabei offensichtliche Artefakte.

—— *Mund-Audio-Synchronisation*: Besonderes Augenmerk sollte man auf konsistente Muster von Ausrichtungsfehlern legen. Natürliche Verzögerungen treten nur sporadisch auf, während die Desynchronisation von Deepfakes ein ständiges Problem darstellt.

### Empfehlungen für die Security-Organisation

CISOs und andere Sicherheitsverantwortliche sollten Verifizierungsprotokolle für sensible Kommunikation erstellen. Eine mehrstufige Authentifizierung sowie Prozesse zu etablieren, die über eine visuelle Bestätigung hinausgehen, ist ebenfalls empfehlenswert.

Gesprächsteilnehmer sollten bei wesentlichen digitalen Konferenzen bestimmte, unvorhersehbare Bewegungen oder Interaktionen anfordern. Und zu guter Letzt sollten CISO & Co. sichere sekundäre Kommunikationskanäle zur Verifizierung bereitstellen.

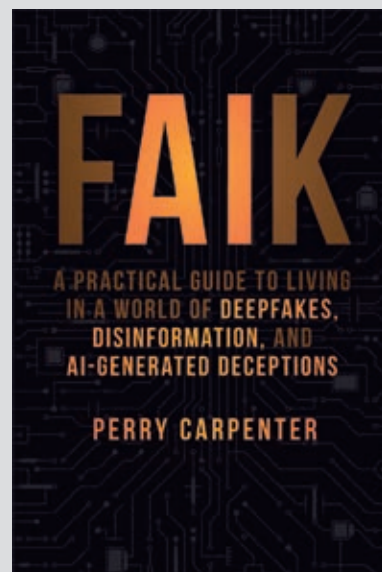
### Fazit

Die aufgeführten Maßnahmen verdeutlichen, dass (und wie) es möglich ist, Deepfakes mit einer natürlichen Skepsis und geistigen Haltung selbst zu erkennen – alles, was dafür nötig ist, sind Übung und Erfahrung. Allerdings – und das zeigen die eingangs erwähnten Fäl-

### Literatur-Tipp: FAIK

In seinem Buch „FAIK: A Practical Guide to Living in a World of Deepfakes, Disinformation, and AI-Generated Deceptions“ [5] zeigt Perry Carpenter nicht nur die Bedrohungslage für Deepfakes auf, sondern gibt auch selbsterprobte Tipps für deren Erkennung. Das Buch bietet einen leicht verständlichen Überblick über generative KI und synthetische Medien. Es zeigt, wie KI-generierte Inhalte als Cyberbedrohung eingesetzt werden. Es streift außerdem die Frage, warum wir auf digitale Manipulationen hereinfliegen und zeigt reale Fallbeispiele sowie deren Folgen auf. Darüber hinaus gibt es dem Leser Strategien an die Hand, um sich selbst und Angehörige vor raffinierten KI-Betrügereien zu schützen.

([www.thisbookisfaik.com](http://www.thisbookisfaik.com))



le nur allzu deutlich – ist es ohne Übung und ohne diese Tipps durchaus möglich, Mitarbeiter und Manager in die Falle tappen zu lassen.

Cyberkriminelle und bösartige Akteure im staatlichen Auftrag nutzen heute (wie jeder andere) verfügbare Technologie, um ihre Ziele zu erreichen – und diese sind oftmals monetärer Natur. Deshalb kann eine erfolgreiche Deepfake-Attacke teuer werden. Trainings für Mitarbeiter von Finanzabteilungen, für Manager und alle, die Kundenkontakt haben oder viel kommunizieren müssen, sind in diesem Licht eine betrachtenswerte Investition.

Der gesamten Security-Community steht eine Zukunft bevor, in der man nicht mehr zwischen Fälschung und Original unterscheiden kann. Mehr und mehr CEOs werden sich damit konfrontiert sehen, dass sie sich zwischen einer intensiven Medienpräsenz und der Gefahr, als Deepfake zu enden, entscheiden müssen.

Wenn das Vertrauen in Bild, Ton und Text grundlegend verloren geht, hat die Menschheit jedoch noch weit größere Probleme als sich mit CEO-Fraud auseinanderzusetzen: Der alte Spruch „das Internet vergisst nichts“, mag zwar so heute nicht mehr haltbar sein – jedoch reichen mit weiteren Technologiesprüngen vielleicht bald bereits kleinere Schnipsel aus, um Deepfakes oder Voice-Fakes zu erstellen.

CISOs und andere Security-Verantwortliche müssen sich auf diese Zukunft vorbereiten und gegebenenfalls schon heute Taskforces für die Erkennung von Deepfakes aufstellen. Spezielle Software, die aktuell bereits verfügbar ist, um Video- oder Ton-Fälschungen zu erkennen, dürfte in Bälde ebenso ausgereift sein wie zukünftige Angriffsmechanismen – und ebenfalls mit KI-Unterstützung arbeiten.

## Literatur

- [1] Presse- und Informationsamt der Bundesregierung, Was ist eigentlich ein Deep Fake?, Experten-Interview mit Oberstleutnant i. G. Aldo Kleemann, Oktober 2023, [www.bundesregierung.de/breg-de/service/archiv/was-sind-deep-fakes-2230226](http://www.bundesregierung.de/breg-de/service/archiv/was-sind-deep-fakes-2230226)
- [2] Monika Dunkel, Der falsche Bill: So wurde der Bayer-Chef Opfer eines Deepfakes, Capital+, Februar 2025, [www.capital.de/35486154.html](http://www.capital.de/35486154.html) (kostenpflichtig)
- [3] Helmut Martin-Jung, Hier spricht der Ferrari-Chef – nicht, Süddeutsche Zeitung, Juli 2024, <https://sz.de/lux.D6i4WqmRwWhL9iT8spKGrG>
- [4] Simon Hurtz, Angestellter überweist 24 Millionen Euro an Betrüger, Süddeutsche Zeitung, Februar 2024, <https://sz.de/1.6344209>
- [5] Perry Carpenter, FAIK, A Practical Guide to Living in a World of Deepfakes, Disinformation, and AI-Generated Deceptions, Wiley, September 2024, ISBN 978-1-394-29988-1, <https://www.wiley.com/en-ie/Edition-p-9781394299898>

Zusammen mit den vorgestellten Maßnahmen der Security-Awareness lassen sich Mitarbeiter schützen und Schäden gleichermaßen von Personen und Unternehmen fernhalten. ■

*Dr. Martin Krämer ist Security Awareness Advocate bei KnowBe4.*

## Werden Sie Fachkraft für IT-Sicherheit!

Kostengünstiges und praxisgerechtes Fernstudium ohne Vorkenntnisse. Vorbereitung auf das **SSCP- und CISSP-Zertifikat**. Ein Beruf mit Zukunft. Beginn jederzeit.

**Teststudium  
ohne Risiko!**

**GRATIS-Infomappe  
gleich anfordern!**



FERNSCHULE WEBER - seit 1959  
Telefon 04487 / 263 - Abt. C99

[www.fernschule-weber.de](http://www.fernschule-weber.de)



**Fernstudium  
Datenschutz-  
beauftragter TÜV**



**Telefon: 04487 / 263**

[www.fernschule-weber.de](http://www.fernschule-weber.de)