Quelle: Александр Марченко/Adobe-Stock

# GENERATIVE KÜNSTLICHE INTELLIGENZ UND IHRE AUSWIRKUNGEN AUF DIE CYBERSICHERHEIT

Impulspapier | Juni 2025

## Zusammenfassung

Von dem Einsatz generativer Künstlicher Intelligenz (KI) werden beträchtliche wirtschaftliche Impulse erwartet. So geht eine McKinsey-Studie[1] von einem weltweiten, wirtschaftlichen Wachstumspotenzial von jährlich zwischen ca. 2 und 4 Billionen US-Dollar aus. Auch für das Gebiet der Cybersicherheit verspricht der kontrollierte Einsatz generativer KI erhebliche Mehrwerte und Chancen[2]. Gleichzeitig gehen mit ihrem Einsatz aber auch erhebliche neue Risiken für die Cybersicherheit einher. So kann die Nutzung generativer KI als Angriffswerkzeug ein Ausgangsunkt für neue und wirkungsvollere Angriffe und damit verbundene ökonomische und gesamtgesellschaftliche Schäden sein. Andererseits kann durch die Nutzung generativer KI der Schutz und die Resilienz von digitalen Systemen substanziell verbessert und Angriffen, auch neuen Angriffsklassen, die durch generative KI automatisiert generiert werden, wirksamer begegnet werden.

Das Papier skizziert sowohl die Risiken als auch insbesondere die Chancen des Einsatzes generativer KI zur Erhöhung der Cybersicherheit. Generative KI könnte in der Fläche zu einer wirksamen Reduktion der klassischen wie auch der neuen Risiken beitragen. Die positiven Potenziale der generativen KI für den Bereich der Cybersicherheit können sich aber nur dann entfalten, wenn die generative KI Vertrauenswürdigkeit („Trustworthy AI“) gewährleistet, das heißt, sie muss robust, privatsphärenerhaltend, transparent, nachvollziehbar und diskriminierungsfrei arbeiten. Ansonsten verwandeln sich vermeintliche Chancen in zusätzliche Risiken.

Die rasante Entwicklung bei generativer KI zeigt, dass die mit sehr hohen Investitionen entwickelten großen, proprietären, generativen Modelle bereits nach sehr kurzer Zeit veraltet sind. Deren Leistungsfähigkeit wird schnell durch neue Modelle, zum Teil mit neuen, ressourceneffizienteren Ansätzen sowie auch durch unabhängige, kostenlose Open-Source-Lösungen übertroffen. Der mit hohen Investitionssummen erzielte Vorsprung der Anbieter generativer KI-Modelle der ersten Stunden erweist sich als flüchtig. Eine nachhaltig wirkende digitale Transformation in Unternehmen und Behörden benötigt deshalb andere Ansätze. Darin liegt eine Chance für Industrie, Forschung und Verwaltung in Deutschland, solche Lösungen bereitzustellen.

Das Papier richtet sich an die Politik und formuliert Empfehlungen, die sowohl dringende Forschungsfragen adressieren als auch auf Umsetzungsthemen eingehen, die teilweise einen Beitrag zur Umsetzung der Anforderungen des EU AI Acts leisten könnten. Dazu gehören Empfehlungen zur Entwicklung und Erprobung von Testverfahren sowie zur Etablierung von vertrauenswürdigen Plattformen, die den sicheren Betrieb KI-basierter Ökosysteme ermöglichen[3]. Wenn Umsetzungsempfehlungen und Referenzimplementierungen schnell bereitgestellt werden, trägt dies zur Rechtssicherheit für KMU, Industrie und den Dienstleistungssektor bei.

## 1. Einführung: Generative KI

Generative KI ist eine Variante der KI, die es ermöglicht, automatisiert Inhalte zu erstellen. Generierte Inhalte können beispielsweise Dokumentationen, Zusammenfassungen und Erklärungen sein, aber auch Bilder oder Programmcode – also Software. Den Generatoren liegen tiefe neuronale Netze (Deep Neural Networks) zugrunde, die auf umfangreichen Corpora (Bücher, Artikel, Webseiten etc.) trainiert werden. Bei den großen Sprachmodellen, den Large Language Models (LLM), werden Milliarden sogenannter Parameter zum Anlernen der Basismodelle, den sogenannten Foundation Models, genutzt, wobei unterschiedliche Trainingstechniken verwendet werden (siehe Abschnitt 2.2). Das Anlernen der Basismodelle mit ihren Milliarden von Parametern ist mit sehr hohen Kosten verbunden und erfordert einen extrem hohen Rechenaufwand sowie enorme Mengen an Strom. Die Anzahl der für das Training benötigten Parameter ist häufig nicht offengelegt. Während man beispielsweise bei der mittlerweile veralteten Version ChatGPT-3 von OpenAI noch von ca. 175 Milliarden (109) Parametern ausgeht, sollen es bei ChatGPT-4 bereits über eine Billion (1012) Parameter gewesen sein. Solche großen Modelle können dann in Ver-

feinerungsschritten für bestimmte Aufgaben wie Spracherzeugung, Generierung von Programmcode oder auch für den Einsatz in bestimmten Anwendungsdomänen speziell zugeschnitten (Fine-Tuning, vgl. 2.2) werden.

Vereinfacht gesagt, arbeitet ein generatives Sprachmodell so, dass neue Texte unter Rückgriff auf das Wissen aus den antrainierten Daten generiert werden. Die Generierung basiert auf der Berechnung von Wahrscheinlichkeiten, den Wortübergangswahrscheinlichkeiten. Das bedeutet, dass das System aus den angelernten Daten ermittelt, was die wahrscheinlichste Antwort auf eine gestellte Frage ist. Daher kommt es immer wieder zu falschen oder gar absurden Antworten, wenn für die gestellte Frage keine verlässliche Datenbasis vorliegt. Umgangssprachlich wird dieses Verhalten der LLM häufig als „Halluzinieren" bezeichnet. Das Halluzinieren ist eine zentrale Herausforderung im Hinblick auf die Vertrauenswürdigkeit von KI. Denn gerade bei KI-basierten Assistenzsystemen können nicht verifizierte Inhalte in den generierten Dokumenten eingebracht werden.

Die mit generativen KI-Modellen verbundenen Fähigkeiten haben bereits heute tiefgreifende Auswirkungen auf viele industrielle Bereiche und auch auf Prozesse in der öffentlichen Verwaltung. Auch der Bereich der Cybersicherheit ist stark betroffen. In Abschnitt 3 geben wir einen Einblick in den wachsenden Einfluss generativer KI auf die Cybersicherheit. Dabei werden sowohl die potenziellen Vorteile als auch die mit der Technologie einhergehenden Risiken skizziert. Aus Sicht der Cybersicherheit ist wichtig zu wissen, welche Information über das jeweilige Modell und die genutzten Trainingsdaten öffentlich zugänglich ist und mit welchen Techniken ein Modell kontrolliert feinabgestimmt trainiert werden kann, zum Beispiel mit individuell gewählten, kuratierten Daten. Deshalb geben wir im folgenden Abschnitt zunächst einen Einblick in bekannte Open- und Closed-Source-Modelle sowie mögliche Trainingstechniken.

## 2. Modell-Klassen und Trainingstechniken

Die aktuelle Landschaft an generativen KI-Modellen ist sowohl von Open-als auch von Closed-Source-Modellen geprägt.

### 2.1 Open- und Closed-Source-Modelle

Mit dem Begriff **Open Source LLM** werden gemeinhin Modelle adressiert, deren Code, Trainingsdaten, Modellarchitektur und Gewichte öffentlich zugänglich sind und unter einer Lizenz veröffentlicht werden, die eine freie Nutzung, Veränderung und Weiterverbreitung erlaubt. Man spricht von **Open Weights LLM**, wenn nur die Gewichte veröffentlicht sind, aber nicht die Trainingsdaten. Manche Open-Weights-Modelle haben auch restriktivere Nutzungslizenzen und sind zum Beispiel nicht für kommerzielle Zwecke nutzbar. Ein Beispiel für ein wirklich offenes Modell wäre Mistral 7B, während Modelle wie LLaMA 2 von Meta zwar Open Weights bieten, aber keine Open-Source-Lizenz im klassischen Sinne haben.

Die von führenden Technologieunternehmen entwickelten Modelle sind meist **Closed Source**. Der Quellcode und auch die Trainingsdaten sind nicht offengelegt. Die Modelle sind kommerzielle Software-Werkzeuge, die häufig spezielle Anwendungsfälle besonders gut unterstützen. So haben beispielsweise GPT-3 und GPT-4 von OpenAI neue Maßstäbe in der Verarbeitung natürlicher Sprache gesetzt, mit bemerkenswerten Fähigkeiten in der Texterzeugung und im Textverständnis. Googles Gemini 2.0 vom Dezember 2024 ist ein weiteres bemerkenswertes Closed-Source-Basismodell, das multimodal ist und dadurch sowohl Texte als auch Bilder und Programmcode verarbeiten kann. Die jüngste Closed-Source-Initiative von Google, Gemma, gewinnt ebenfalls an Bedeutung und bietet speziell auf die Konversation zugeschnittene Modelle, sodass damit insbesondere KI-basierte Assistenzfunktionen und Chat-Bots unterstützt werden.

Mittlerweile gibt es aber auch als **Open Source** sehr leistungsstarke Angebote. Eine zunehmend breitere Akzeptanz finden die verschiedenen Varianten des Modells Mistral des französischen Unternehmens Mistral AI, das mit Mistral NeMo auch ein Open-Source-Modell anbietet. Eine breite Zugänglichkeit zu Open-Source-Modellen wird durch Plattformen wie Hugging Face erreicht. Diese erlaubt es Nutzenden, mit verschiedenen Modellen zu experimentieren und Modelle gezielt für eine breite Palette von generativen KI-Anwendungen anzupassen. Die Transformer-Bibliothek von Hugging Face hat zusammen mit Open-Source-Modellen wie GPT-Neo von EleutherAI, Bloom von BigScience und Teuken-7B der Fraunhofer-Gesellschaft die Zugänglichkeit von LLM erheblich verbessert und damit beim Einsatz generativer KI schnelle Fortschritte in der Breite ermöglicht. DeepSeek ist ein chinesisches Unternehmen mit seit Ende Januar 2025 hoher Medienpräsenz. Eines seiner Open-Source-Modelle, DeepSeek-V3, bietet ebenfalls eine sehr große Leistungsfähigkeit, die mit führenden Closed-Source-Modellen vergleichbar ist. Derzeit ist noch unklar, wieviel Rechenleistung und Aufwand tatsächlich für das Trainieren der Modelle eingesetzt wurde. Das Modell basiert auf einer sogenannten Mixture-of-Experts(MoE)-Architektur. Das bedeutet, dass Anfragen durch gezielt ausgewählte Expertenmodelle beantwortet werden. Da intransparent ist, mit welchen Daten das Modell trainiert wurde und die Ausgaben zudem auch unter chinesischer Zensur stehen, ist die Nutzung des Modells (Stand Februar 2025) in Regierungsbehörden durch einige Länder wie Australien und Italien be-

reits verboten worden. In den USA diskutiert man ebenfalls über ein Verbot. Trotz der vielen offenen Fragen, und einer sicherlich auch angebrachten großen Zurückhaltung, ist der gewählte neue Ansatz interessant und könnte auch für europäische beziehungsweise deutsche Entwicklungen eine Blaupause sein – natürlich mit kuratierten Daten und ohne Zensur.

Insgesamt beflügelt die Bereitstellung von Open- und Closed-Source-Modellen die Nutzung generativer KI-Technologien. Auch wenn in den Anfängen die US-amerikanische Vorherrschaft bei großen Sprachmodellen unübersehbar war und teilweise noch ist (unter anderem OpenAI, Google, Microsoft, Meta), so gibt es auch auf der **nationalen und europäischen Ebene** bereits seit einigen Jahren und zunehmend verstärkt Ansätze zur Entwicklung von LLM. Zu nennen sind hier insbesondere das französische Modell Mistral, oder auch das im November 2024 veröffentlichte Modell Teuken-7B des Projekts OpenGPT-X, das unter deutscher Leitung entstanden ist, oder auch Bloom, ein Sprachmodell der kollaborativen BigScience-Initiative. Weitere europäische Projekte, wie TrustLLM unter Leitung des Forschungszentrums Jülich und OpenEuroLLM, sind ebenfalls bereits gestartet. Daneben gibt es bekannte Ansätze wie die weiterentwickelten, vollkommen offenen Pharia-1-Modelle von Aleph Alpha und für den Bereich der Übersetzungen DeepL; beide Systeme stammen von deutschen Firmen aus Heidelberg beziehungsweise Köln. Daneben ist auch die europäische Current-AI-Initiative zu nennen, die auf eine dezentrale Verarbeitung setzt. Es gibt also bereits erhebliche Anstrengungen auf der europäischen und auch auf der nationalen Ebene, alternative LLM bereitzustellen. Diese Entwicklungen werden durch öffentliche Mittel substanziell gefördert. Zudem werden in Deutschland auch für den Aufbau von Rechenkapazitäten bereits erhebliche öffentliche Mittel bereitgestellt. **Deutschland und Europa sind durchaus in der Lage, große Sprachmodelle zu entwickeln.** Um nachhaltig einen Mehrwert für Unternehmen in Deutschland zu bieten, scheinen jedoch allgemein einsetzbare, aufwändig trainierte und schnell alternde große Sprachmodelle eher ungeeignet. Das Papier schlägt deshalb einen ergänzenden Weg vor.

### 2.2 Trainingstechniken

Ansätze für das gezielte Training von generativen KI-Modellen unterscheiden sich darin, welche Ressourcen genutzt und welche spezifischen Ziele verfolgt werden. Eine gängige Methode ist das sogenannte **Fine-Tuning**, bei der ein vorab trainiertes Modell an bestimmte Aufgaben oder Daten angepasst wird. Dies ermöglicht eine schnelle Anpassung an bestimmte Anforderungen und spart Rechenressourcen, da das Modell nicht von Grund auf neu trainiert werden muss. Eine weitere Technik ist das **Transfer-Lernen**, bei dem Modelle auf der Grundlage von bereits vorhandenem Wissen für eine verwandte Aufgabe oder einen Bereich weiter trainiert werden, wobei nur ein Teil der Parameter angepasst wird. Weitere Alternativen sind das **Prompt-Engineering**, bei dem die Eingaben in die Modelle verfeinert werden, um sie auf die gewünschten Ergebnisse zu lenken, und das **„Few-Shot Learning“**, bei dem die Modelle mit einer minimalen Anzahl von Beispielen trainiert werden und sich auf ihre Fähigkeit zur Verallgemeinerung aus wenigen Daten verlassen. Die Fähigkeit, mit wenigen Beispielen angelernt werden zu können, ist Fluch und Segen zugleich, da dies natürlich auch ein Einfallstor für eine gezielte Manipulation der Modelle ist.

Von besonderer Bedeutung insbesondere auch im Hinblick auf Transparenz, Nachvollziehbarkeit und Kontrollierbarkeit ist der Ansatz der **Retrieval-Augmented Generation (RAG)**. Bei RAG werden generative Modelle mit einer externen Datenbank kombiniert, sodass das Modell während des Generierungsprozesses auf eine Datenbank oder eine Wissensbasis zugreifen kann. Dies verbessert die Fähigkeit des Modells, genauere, aktuellere und kontextuell relevante Antworten zu generieren, indem es Informationen in Echtzeit abruft, ohne dass das Modell für jede neue Information neu trainiert werden muss. Über die externe Wissensbasis ist es möglich, verlässliche, geprüfte Daten und Kontextinformationen in den Trainingsprozess einzuspeisen und damit nicht nur die Transparenz zu erhöhen, sondern auch die LLM in ihren Ausgaben kontrollierbarer und nachvollziehbarer zu gestalten. Halluzinieren kann dadurch zwar nicht gänzlich unterbunden, jedoch deutlich verringert werden. Schließlich sei angemerkt, dass neueste Forschungsarbeiten sogar operativ im Parameterraum der Modelle im Inferenzbetrieb ansetzen, um ein gewünschtes Verhalten zu erzwingen.

Das dedizierte Trainieren von LLM mit kontrollierten und effizienten Trainingsmethoden und mit kontrollierten, kuratierten Daten, wie beispielsweise aus dem Verwaltungsumfeld, aus dem Betrieb von (kritischen) Infrastrukturen, oder aus der Fertigungsindustrie, eröffnet die Chance, schnell, ressourcenschonend und kontrolliert dedizierte LLM für spezifische Einsatzbereiche zu entwickeln. **Es ist deshalb essenziell, dass die dafür erforderliche vertrauenswürdige Schlüsseltechnologie im Bereich Trusted AI mit hoher Innovations- und Transfergeschwindigkeit in Deutschland erforscht und entwickelt wird.** Hierzu gehören unter anderem neue Verfahren und Methoden, um die Robustheit und Vertrauenswürdigkeit generativer Modelle zu testen sowie die Erforschung und Entwicklung von Zertifizierungsschemata, um die Güte der Modelle vergleichbar zu attestieren.

## 3. Risiken und Chancen generativer KI für Cybersicherheit

Im Folgenden werden einige der wichtigsten Risiken, aber auch Chancen aufgezeigt.

### 3.1 Risiken für die Cybersicherheit

Bereits heute sind generative KI-Verfahren und -Systeme integraler Bestandteil von Systemlandschaften, sei es bei Consumer-Produkten, in der Office-IT und zunehmend auch in der Operational-IT (OT), wie industriellen Automatisierungsanlagen. Sie sind damit Bestandteil von IT-Systemen und können selbst das Ziel gezielter Angriffe sein. Bereits heute basieren vielfältige, automatisiert umgesetzte Aktionen und Reaktionen auf Empfehlungen und Ergebnissen generativer KI-Verfahren. Der potenzielle Schaden und die Auswirkungen, die gezielt manipulierte generative KI-Verfahren auf den Betrieb von IT-Systemen haben können, sind somit bereits jetzt absehbar und sehr kritisch. Im Folgenden gehen wir jedoch auf diese Klasse von Risiken nicht ein, da Fragen der gezielten Manipulation von KI-Verfahren bereits seit längerer Zeit in Forschung und Entwicklung bearbeitet werden.

Durch generative KI ergeben sich aber auch **neue oder verstärkte Risiken**, wie das Risiko des unbeabsichtigten Informationsabflusses durch Anfragen an KI-Chatbots. Im Folgenden werden mögliche Risiken grob in drei Kategorien aufgeteilt. In der ersten Kategorie fassen wir Risiken zusammen, die einzelne Nutzende betreffen und darüber auch die Organisation, wenn die Person im beruflichen Kontext handelt. Die zweite Kategorie adressiert neue beziehungsweise verstärkte Risiken, die sich aus der automatisierten Generierung von konkreten Angriffskampagnen ergeben. Die dritte Kategorie beschäftigt sich mit KI-basierter Angriffsassistenz, die die Generierung neuer, komplexer Angriffe unterstützen beziehungsweise sogar erst ermöglichen. Die beiden letzten Kategorien sprechen damit Risiken an, die für die Cybersicherheit im Großen eine zentrale Rolle spielen beziehungsweise zukünftig vermehrt spielen werden.

*(1) Individuelle Nutzungsrisiken:*

- Zu den besonderen Risiken, die mit LLM in Verbindung gebracht werden[4], gehören die gezielte Manipulation der Eingaben in LLM, bekannt als Prompt Injection. Durch die Interaktion mit LLM kann es zu einer unbeabsichtigten Weitergabe von vertraulicher Information kommen, etwa wenn Mitarbeitende in Anfragen Firmeninterna preisgeben.
- Aber auch über die Ausgaben von LLM kann es zu einer unbeabsichtigten Offenlegung vertraulicher Daten kommen, wenn beispielsweise aus den Antworten des Sprachmodells Rückschlüsse auf personenbezogene Daten, mit denen das Modell trainiert wurde, möglich sind, bekannt als Inference Attack.
- Generell gelten für Ausgaben von LLM analoge Risiken wie sie aus dem Bereich der Web-Anwendungen schon bekannt sind. So kann die ungeprüfte Weitergabe und Weiterverarbeitung von LLM-Ausgaben dazu führen, dass Schadcode in Unternehmensinfrastrukturen eingebracht und zur Ausführung kommen kann, mit den bekannten Risiken, wie des unautorisierten Datenabflusses, der Manipulation von Daten und generell der Störung von Abläufen.
- Die Trainingsdaten spielen wie bei jeder KI eine herausgehobene Rolle. Hier setzen eine ganze Reihe von Risiken an. Als erstes ist das gezielte Manipulieren von Trainingsdaten zu nennen, bekannt als Training Data Poisoning, um das Verhalten des LLM zu manipulieren. Risiken können von sehr unterschiedlichen Komponenten, die im Lebenszyklus eines LLM eine Rolle spielen, hervorgerufen werden. Hierzu gehören etwa Datensätze von Drittanbietern oder auch vortrainierte Modelle und Plug-ins, durch die es zu unerwünschten Zugriffen kommen kann.
- Auch fehlerhafte Ausgaben der generativen KI können zu Sicherheitsrisiken führen. So könnten falsche Antworten eines LLM, wie beispielsweise fehlerhafter Programmcode, wenn sie ungeprüft vom Nutzer übernommen und in die eigene Software integriert werden, zu neuen Sicherheitsrisiken führen.

*(2) KI-generierte Klassen von Angriffen*

- Mit generativer KI kann anpassungsfähiger Schadcode (Malware) automatisch generiert werden, ohne dass Angreifende über spezifische Kenntnisse verfügen müssen. Diese Anpassungsfähigkeit könnte zu einer neuen Generation von polymorphen Viren führen, das heißt Malware, die ihre Eigenschaften dynamisch verändern kann, um herkömmliche Abwehrmaßnahmen zu umgehen.
- Mit generativer KI und unter Nutzung offen zugänglicher Daten können Deep Fakes generiert und autonom breitflächige Phishing-Kampagnen gestartet werden.
- Generative KI kann auch zur Erstellung und Verbreitung von Fake News und orchestrierten Desinformationskampagnen eingesetzt werden. Die Risiken gehen über die Schädigung von Einzelpersonen hinaus und reichen bis zu umfassenderen politischen und gesellschaftlichen Verwerfungen. Diese Kampagnen können das Vertrauen der Öffentlichkeit erschüttern, Meinungen manipulieren und damit in demokratische Prozesse eingreifen.
- Die Fähigkeiten autonomer KI-Agenten verbessern sich rasant. Mithilfe von Open Source Intelligence (OSINT) und anderen Datenquellen können diese Agenten automatisch Schwachstellen in Systemen ausfindig machen.

Es ist zu erwarten, dass auf dieser Basis neue Angriffstechniken entwickelt werden. So könnte eine Hierarchie spezialisierter Agenten zum Einsatz gebracht werden, wobei jeder Agent für bestimmte Aufgaben wie Aufklärung, Schwachstellenanalyse oder die Entwicklung von Angriffsvektoren spezialisiert trainiert ist. Diese Angriffsstrategien können von der gezielten Ausnutzung der IT-Infrastruktur mit technischen Mitteln bis hin zum Social Engineering und der Manipulation von Geschäftsprozessen reichen. Die Automatisierung und Skalierbarkeit dieses Ansatzes erhöhen die Risiken für Unternehmen oder Behörden erheblich.

- Herkömmliche Sicherheitsanalyse-Frameworks zur Unterstützung von Sicherheitsverantwortlichen, wie zum Beispiel Metasploit für die Durchführung von Penetrationstests, sind seit langem ein fester Bestandteil von Cybersicherheitsmaßnahmen. Die Integration großer LLM in diese Frameworks könnte zu einer neuen Klasse hybrider Angriffe führen. KI-Agenten könnten mehrere Werkzeuge und Frameworks gleichzeitig orchestrieren und ihre Angriffsstrategien dynamisch anpassen, um erfolgreiche Angriffe durchzuführen. Diese Verschmelzung von KI mit herkömmlichen Sicherheitswerkzeugen erhöht sowohl die Effizienz als auch die Effektivität potenzieller Cyberangriffe.

*(3) KI-unterstützte Angriffsvorbereitung*

- Speziell auf die Erkennung von Schwachstellen trainierte LLM, die für Sicherheitsverantwortliche eine große Hilfe sein könnten, helfen aber auch Angreifenden bei der automatisierten Identifikation möglicher Schwachstellen in Zielsystemen. Mit Unterstützung generativer KI können Angreifende schneller wirkungsvolle Angriffe entwickeln, wodurch sich sowohl die Häufigkeit als auch die Auswirkungen solcher Angriffe drastisch erhöhen werden.
- Fortgeschrittene LLM sind besonders gut in der Lage, komplexe, unstrukturierte oder auch strukturierte Daten aus verteilten Quellen zu verarbeiten und in Form von zum Beispiel Zusammenfassungen, Empfehlungen oder auch auszuführendem Programmcode bereitzustellen. Diese Fähigkeit kann auch für die Angriffsvorbereitung gezielt genutzt werden. Durch die Kombination von OSINT mit anderen frei verfügbaren Datenquellen können sich Angreifende automatisiert ein umfassendes und verwertbares Lagebild ihrer Ziele verschaffen. Dies erleichtert es, komplexe, hochwirksame Cyberangriffe zu planen und durchzuführen, selbst wenn die Angreifenden über sehr wenig technische Expertise verfügen.

**Über generative KI werden bestehende Risiken massiv verstärkt und es kommen neue, derzeit nur zu erahnende Risiken hinzu.** Diese reichen von der autonomen Erstellung von Lagebildern für Angreifende über automatisiert generierte, wirksame Angriffskampagnen bis hin zu groß angelegtem Social Engineering und Desinformationskampagnen.

### 3.2 Chancen für die Cybersicherheit

Um die mit generativer KI einhergehenden zusätzlichen Risiken wirksam abzuschwächen, müssen Unternehmen und Behörden sowohl in geeignete Sicherheitstechnologie investieren, als auch ein umfassendes Risiko- und Schwachstellenmanagement etablieren und dessen Wirksamkeit kontinuierlich evaluieren. Diese Forderungen sind nicht neu, deren Umsetzung in Unternehmen und Behörden ist aber spätestens jetzt, angesichts der deutlich erhöhten Risikolage durch generative KI, unvermeidlich.

Hier setzen aber auch Chancen an, die in der Nutzung generativer KI liegen. Mit dem gezielten Einsatz von generativer KI kann es nicht nur gelingen, klassische Risikolagen abzumildern, sondern auch den neuen Risikoklassen, die durch generative KI wie oben beschrieben entstehen, wirksam zu begegnen. In diesem Feld sind aber noch erhebliche Forschungs- und Entwicklungsanstrengungen notwendig. Bereits jetzt gibt es sehr ermutigende Ansätze, die zeigen, welche Lösungsansätze effizient, wirksam und nachhaltig sein werden. Damit könnte es sogar mittel- bis langfristig möglich sein, Geschäftsmodelle von Angreifenden zu zerstören, da die Kosten für die Erstellung von wirksamen Angriffskampagnen den zu erwartenden Gewinn deutlich übersteigen. Es könnte somit sogar die Chance bestehen, breiten Kreisen von Angreifenden die Geschäftsgrundlagen zu entziehen.

Die Anwendung von **KI zur Erhöhung der Cybersicherheit** bietet Vorteile in verschiedenen Gebieten. Nachfolgend sind drei wichtige Anwendungsbereiche skizziert: (1) Generative KI kann gezielt proaktives Handeln unterstützen, sodass Systeme resilienter gegen klassische und KI-generierte Angriffe sind, (2) generative KI kann Sicherheitsverantwortlichen bei der Erfüllung von operativen Cybersicherheitsaufgaben assistieren und (3) generative KI kann zur Verbesserung der Qualität von Software oder auch des Systemdesigns genutzt werden.

*(1) Proaktive Erhöhung der Cybersicherheit auch gegen KI-basierte Angriffe*

- Generative KI kann sehr große Datenmengen mit beispielloser Geschwindigkeit verarbeiten und ermöglicht so die Erkennung von und Reaktion auf Bedrohungen

in Echtzeit (zum Beispiel Anomalieerkennungs- oder Intrusion-Detection-Systeme). KI-Werkzeuge können umfassende Einblicke in laufende Angriffe bieten, indem sie Kausalketten identifizieren und die Hauptursachen, zum Beispiel ausnutzbare Schwachstellen, für den Angriff aufzeigen, was eine schnellere und präzisere Reaktion ermöglicht. Damit könnte mit KI-basierten Abwehrmaßnahmen wirksam und automatisiert auch KI-generierten Angriffskampagnen begegnet werden.

- Durch die kontinuierliche Datenanalyse können mögliche Angriffswege prognostiziert und Sicherheitsverantwortliche darin unterstützt werden, die Schutzmaßnahmen vorbeugend dort zu verbessern, wo mit hoher Wahrscheinlichkeit Angriffe zu erwarten sein werden. Dies ist eine Form der Predictive Security AI und ist dort zu erwarten, wo generative KI eigenständig Schwachstellen identifizieren und selbstständig Angriffe und Gegenmaßnahmen formulieren kann. Derzeit sind solche Nutzungen von KI zur Angriffsabwehr jedoch noch nicht im Einsatz.
- Die Erkennung und das automatisierte Einleiten von Reaktionen können auch ohne ein menschliches Eingreifen erfolgen, was die Reaktionszeiten erheblich verkürzt. In diesem vollkommen autonomen Ansatz „arbeitet" man mit generativer KI gegen Angriffe, die durch generative KI erfolgen. Dies setzt aber voraus, dass die automatisiert eingeleiteten Maßnahmen nachvollziehbar sind und die Entscheidungen auf kuratierten Daten von hoher Qualität und Vertrauenswürdigkeit beruhen. Zudem ist die Fähigkeit von KI-Systemen, kausale Zusammenhänge über einfache Korrelationen hinaus zu erkennen, (noch) nicht hinreichend entwickelt und muss grundlegend erforscht werden (siehe dazu die Handlungsempfehlungen).

*(2) Assistenz bei der Erfüllung von Cybersicherheitsaufgaben*

- Generative KI kann genutzt werden, um Routineaufgaben im Bereich des Cybersicherheitsmanagements zu automatisieren. Beispiele für solche Aufgaben sind die Analyse von Log-Dateien, um beispielsweise auffälliges Nutzungsverhalten zu detektieren und zu dokumentieren. Damit leistet generative KI auch einen Beitrag, um dem Fachkräftemangel in der Cybersicherheit zu begegnen. Sie lässt sich zudem dafür einsetzen, automatisierte, KI-generierte Standard-Angriffskampagnen abzuwehren, was die Hürde für Angreifende deutlich erhöht.
- Ein effektives Schwachstellenmanagement ist von entscheidender Bedeutung. Dies findet auch seinen Niederschlag in gesetzlichen Vorgaben wie dem Cyber Resilience Act (CRA). Generative KI kann dabei helfen, automatisiert Schwachstellen zu identifizieren, zu verfolgen und zu beheben, um die Resilienz der Systeme zu erhöhen und deren Sicherheit zu gewährleisten. Da KI-basierte Angriffsassistenzwerkzeuge auf den allgemein bekannten Schwachstellen aufsetzen, kann damit auch großen Klassen automatisiert generierter Angriffe proaktiv ein Riegel vorgeschoben werden.
- Generative KI-Systeme können die Umsetzung von Compliance-Prozessen unterstützen, die durch diverse Vorgaben wie Datenschutz-Grundverordnung (DSGVO), die überarbeitete EU-Cybersicherheitsrichtlinie NIS2 oder den CRA gefordert werden. Generative KI kann durch die Automatisierung von Compliance-Tests helfen, kontinuierlich die Einhaltung dieser Vorgaben zu prüfen und die Prozesse und Ergebnisse automatisiert zu dokumentieren. Die Kombination von KI mit Frameworks wie der Open Security Controls Assessment Language (OSCAL) von NIST und die Integration der KI in das Common Security Advisory Framework (CSAF) des Bundesamts für Sicherheit in der Informationstechnik (BSI) kann die Effizienz und Genauigkeit der Compliance-Bemühungen verbessern.

Offensichtlich kann aber ein Assistenzsystem basierend auf generativer KI auch wieder zu einem erheblichen Risiko werden. Wiederum wird deutlich, dass der **Einsatz generativer KI gerade für sicherheitskritische Aktivitäten besondere Anforderungen an die Vertrauenswürdigkeit dieser Systeme stellt**. So könnte ein manipuliertes System gezielt fehlerhafte Sicherheitskonfigurationen empfehlen (oder einfach nur herbei halluzinieren). Zudem könnten Ergebnisse von regelmäßig durchgeführten Sicherheitsscans, die wertvolle Informationen über potenzielle Schwachstellen beinhalten, in die Hände von Angreifern gelangen.

*(3) Verbesserung der Sicherheitsqualität*

- IT-basierte Systeme besitzen einen sehr hohen Anteil an Software, die nicht mehr aus einer Hand entwickelt wird, sondern aus verschiedenen Quellen stammt – man spricht hier auch von der Software Supply Chain. Häufig stehen die Programme nur in ausführbarem Binärcode zur Verfügung, sodass Unternehmen nicht prüfen können, ob der Programmcode bekannte Sicherheitsschwachstellen oder gar Schadcode und Hintertüren enthält. Mit Methoden der generativen KI kann die Analyse von Software-Artefakten in Bezug auf die Sicherheit teilautomatisiert unterstützt werden, um beispielsweise anhand von Ähnlichkeitssuche bekannte und dokumentierte Schwachstellen im Programmcode zu finden und diese Analyse auch zu dokumentieren. Damit können Vorgaben zur Gewährleistung der Sicherheit in der Software Supply Chain, wie sie mit NIS2 oder dem CRA auf Unternehmen zukommen, teilautomatisiert unterstützt werden.

- Generative KI kann auch die Softwareentwicklung verändern, indem mit KI-Unterstützung Schwachstellen bereits in den frühen Phasen des Designs und der Programmierung identifiziert und von der KI vorgeschlagene sichere Alternativen genutzt werden. In Kombination mit klassischen Code-Analysen kann generative KI automatisiert nicht nur bekannte Sicherheitslücken erkennen, sondern auch potenzielle Schwachstellen aufdecken, die mit herkömmlichen Methoden nicht erkannt werden können. Zudem kann generative KI sicherheitsoptimierten Code vorschlagen, der bewährte Best Practices berücksichtigt und potenzielle Angriffsvektoren reduziert. Darüber hinaus ermöglicht sie eine kontinuierliche Verbesserung der Codequalität durch die Integration in Entwicklungsumgebungen (Integrated Development Environments, IDEs) und in Continuous-Integration-/Continuous-Deployment-Pipelines (CI/CD-Pipelines), wodurch Sicherheitsprobleme bereits vor der Bereitstellung des Codes behoben werden können. Generative KI hat das Potenzial, die Softwaresicherheit substanziell zu verbessern und Entwicklungsprozesse sowie Softwaretests effizienter und sicherer zu gestalten.

**Generative KI bietet also ein breites Spektrum an Möglichkeiten, die Cybersicherheit zu stärken und Risiken zu mindern.** Von der Automatisierung von Routineprozessen bis hin zur Verbesserung des Situationsbewusstseins und der Einhaltung von Vorschriften. Durch die Integration dieser Technologien in ihre Cybersicherheitsstrategien können Unternehmen und Behörden eine robustere und widerstandsfähigere Verteidigung gegen die sich beständig weiterentwickelnden Bedrohungen gewährleisten.

Jedoch gilt auch hier wieder, dass der Einsatz nicht vertrauenswürdiger generativer KI die Chancen in potenzielle Risiken verwandelt. **Die Entwicklung und Bereitstellung vertrauenswürdiger KI-Systeme ist somit eine unabdingbare Notwendigkeit, um die Chancen der generativen KI für die Cybersicherheit nutzbar zu machen.**

## 4. Empfehlungen an die Politik

Die rasante Entwicklung bei generativer KI zeigt, dass die mit sehr hohen Investitionen entwickelten großen, proprietären, generativen Modelle bereits nach sehr kurzer Zeit – häufig schon nach wenigen Monaten – veraltet sind. Deren Leistungsfähigkeit wird schnell durch neue Modelle übertroffen, teils mit ressourceneffizienteren Ansätzen, teils durch unabhängige, kostenlose Open-Source-Lösungen. Der mit hohen Investitionssummen erzielte Vorsprung der Anbieter generativer KI-Modelle der ersten Stunden erfordert konsequente weitere hohe Investitionen. Experten sind sich einig, dass die Entwicklung von domänenspezifischen Modelle, die dediziert kontrolliert und weiter entwickelt werden, ein zielführender Weg ist. **Hier liegt die Chance der deutschen Industrie, Forschung und Verwaltung, solche Lösungen bereitzustellen.**

Forderungen nach Bürokratieabbau sind in aller Munde. Viele bürokratische Prozesse sind jedoch sehr sinnvoll, verursachen aber häufig auf allen Seiten einen sehr hohen manuellen Aufwand. Mit einer Initiative, die durch Nutzung generativer KI einen **Verwaltungssprung** herbeiführt, könnte ein entscheidender Schritt gemacht werden. Dem Einsatz generativer KI stehen jedoch sehr häufig Sicherheitsbedenken, insbesondere auch Datenschutzbedenken, entgegen. Hier setzt eine der Empfehlungen an. Mit der konkreten Umsetzung einer sicheren Trainings- und Betriebsumgebung für LLM, die dediziert auf Prozesse der Verwaltung zugeschnitten ist, könnten der Nutzen von generativer KI aufgezeigt und mögliche Risiken nachweislich vermieden werden. Mit einem solchen Ansatz, der belegt, wie Verwaltungsorgane ein vertrauenswürdiges KI-Ökosystem etablieren und sicher, nachvollziehbar und souverän betreiben können, könnte eine Blaupause geschaffen werden, die von Unternehmen, Behörden oder auch staatlich geförderten Einrichtungen für ihr jeweiliges Einsatzumfeld adaptiert werden kann. Es soll hierbei nicht darum gehen, die Machbarkeit der KI-Einführung anhand eines einzelnen Anwendungsfalls oder einer spezialisierten Fachlösung zu demonstrieren, denn davon gibt es bereits erfreulich viele. Stattdessen ist es entscheidend, eine Referenzimplementierung für Plattformen zu entwickeln, die als einfach auszurollende vertrauenswürdige Basis für eine Vielzahl unterschiedlicher Anwendungsfälle für KI dienen kann.

Nachfolgend werden Empfehlungen an die Adresse der Politik formuliert, die sowohl dringende Forschungsfragen adressieren, als auch auf Umsetzungsthemen eingehen, die teilweise einen Beitrag zur Umsetzung der Anforderungen des EU AI Acts leisten könnten. Nach Artikel 57 Absatz 1 Satz 1 der Verordnung werden die Mitgliedstaaten verpflichtet, mindestens ein KI-Reallabor (Regulatory Sandbox) auf nationaler Ebene ggf. auch gemeinsam einzurichten, das bis zum 2. August 2026 einsatzbereit sein muss. Derzeit ist geplant, die Bundesnetzagentur mit der Umsetzung zu beauftragen. Die unten ausgeführten Empfehlungen zur Entwicklung und Erprobung von Testverfahren, aber auch die Empfehlungen zur Etablierung von vertrauenswürdigen Plattformen zum sicheren Betrieb von KI-basierten Ökosystemen könnten einen Beitrag zu den Aufgaben dieser Reallabore leisten. Die Erkenntnisse aus bereits bestehenden Aktivitäten, wie zum Beispiel dem KIPITZ für Behörden der Bundesverwal-

tung, müssen hierbei natürlich berücksichtigt werden. Die schnelle Bereitstellung von Umsetzungsempfehlungen und Referenzimplementierungen würde zudem einen Beitrag zur Rechtssicherheit für KMU, Industrie und den Dienstleistungssektor leisten.

**Empfehlung 1: Erforschen neuer Ansätze für generative KI-Ökosysteme**

Mit der Entwicklung dedizierter, kontrollierbarer Modelle, die mit dem speziellen Domänenwissen von einzelnen Unternehmen, Branchen oder Verwaltungen trainiert und auf domänenspezifische Aufgabenerfüllung hin ausgerichtet werden, die souverän und flexibel auf Änderungen reagieren oder auch neue Ansätze, wie agentenbasierte KI, aufgreifen können, können Deutschland und Europa ihren eigenen Weg gehen: schnell, ressourcensparend, qualitativ hochwertig, hocheffizient und verlässlich. Investitionen in Forschung und Entwicklung, die darauf abzielen, neue Ansätze für generative KI-Ökosysteme zu erforschen, sind dringend erforderlich. Diese Ansätze können strategisch dazu genutzt werden, sowohl die Abhängigkeit von proprietären Lösungen zu reduzieren, als auch regulatorische und ethische Rahmenbedingungen in Europa konsequent umzusetzen, um so souveräne und vertrauenswürdige KI-Ökosysteme zu etablieren. Gerade für Deutschland und Europa ergeben sich klare wirtschaftliche Chancen durch maßgeschneiderte KI-Lösungen für KMU, Industrie und Verwaltung, die dringend benötigt werden.

i. Benötigt werden **neue Architektur-Ansätze** und vor allem **innovative Trainings- und Inferenz-Methodologien**, sodass Modelle kosteneffizient entwickelt sowie nachhaltig (im Sinne des Energieverbrauchs) und mit geringen Ressourcen, vertrauenswürdig und souverän betrieben werden können. Dies gilt besonders für Systeme mit verteilten „Computer"-Ressourcen, die aktuell nur unter sehr limitierten Bedingungen überhaupt betrieben werden können. Dazu gehört vor allem die Schaffung von Schnittstellen und Algorithmen für ein optimiertes und robustes Training dieser Systeme. Hier sollte eine umfassende Lösung erforscht und entwickelt werden, die vorhandene Trainingsparadigmen komplett „neu denkt", und mittelfristig zentrierte Ansätze ersetzt. Gleichzeitig müssen in diesen neuen Frameworks die Anforderungen des EU AI Acts „von Grund auf" mitgedacht werden, das heißt, es müssen Lösungen in Form von Architekturen und Algorithmen entwickelt werden, die Robustheit, Fairness und Privacy by Design nachweislich umsetzen. Die neuen KI-Ökosysteme müssen architekturell so vorbereitet sein, dass sie schnell und kontrolliert an neue Anforderungen angepasst, aber auch schnell auf gänzlich neue, innovative Architekturen umgestellt werden können.
ii. Es sollte gezielt in die Forschung investiert werden, welche die **theoretischen Grundlagen** der inhärenten Mechanismen von LLM erforscht und damit kontrollierbar macht. Aktuelle Forschung fokussiert auf Erklärungen a posteriori, die den Trainingsprozess ausblenden. Neue Ansätze sind erforderlich, um unter anderem Effekte wie das Halluzinieren wirksamer zu unterbinden. Diese werden derzeit typischerweise mit Prompting-Methoden und RAG-Techniken adressiert, die aber noch zu kurz greifen. Gleichzeitig muss auch die Entscheidungs- und Beurteilungsfähigkeit (das „Reasoning") von LLM signifikant verbessert werden, um deutlich komplexere Aufgaben bewältigen zu können. Dies beinhaltet kausale Zusammenhänge besser zu erkennen und für die Lösungsstrategien zu nutzen. Eine große Chance besteht für Deutschland und Europa darin, KI-Agenten basierend auf kleineren KI-Systemen und deren Interaktion maßgeschneidert zu entwickeln und auch entsprechende Schnittstellen zu schaffen. Der Einsatz generativer KI bei der Identifikation von Software-Schwachstellen durch statische Analyse, der derzeit nachweislich begrenzt ist, wird beispielsweise dadurch unmittelbar beeinflusst.

**Empfehlung 2: Entwicklung vertrauenswürdiger, generativer Cybersicherheitslösungen**

Wie in Abschnitt 3.2 „Chancen für die Cybersicherheit" ausgeführt, kann vertrauenswürdige generative KI substanzielle Beiträge leisten, die Cybersicherheit zu verbessern. Viele der skizzierten KI-basierten Werkzeuge zur Unterstützung der Cybersicherheit erfordern jedoch derzeit noch weitere Forschungs- und Entwicklungsanstrengungen. Bei der Entwicklung von Softwarelösungen, die eine direkte Auswirkung auf den Sicherheitszustand von Einzelpersonen, Unternehmen, aber auch Behörden haben, sollte dringend in die Entwicklung dedizierter, vertrauenswürdiger generativer Cybersicherheitslösungen investiert werden. Damit wird es möglich, die Abhängigkeit von Drittanbietern zu reduzieren und souverän kontrollierbare Lösungen am Markt zu platzieren. In allen drei der in Abschnitt 3.2 skizzierten Anwendungsbereiche[5] wird weitere Forschungs- und Entwicklungsarbeit benötigt. Dazu gehört die Entwicklung KI-basierter Assistenzsysteme, die Sicherheitsverantwortliche darin unterstützen, proaktiv Systeme resilienter gegen klassische und zukünftige KI-generierte Angriffe zu entwickeln, zu konfigurieren und über deren Lebenszeit zu betreiben. Es werden fortgeschrittene, verlässliche Assistenzsysteme benötigt, die Sicherheitsverantwortliche bei den operativen Cybersicherheitsaufgaben unterstützen. Schließlich sind noch weitere Forschungsanstrengungen erforderlich, um generative KI-

Verfahren zur Analyse von großen Softwarepaketen automatisiert und mit hoher Verlässlichkeit einzusetzen.

**Empfehlung 3: Neue Testverfahren und Zertifizierungsschemata**

Es werden neue Testverfahren und -umgebungen erforderlich, um auch generative KI-Systeme hinsichtlich ihrer Vertrauenswürdigkeit und Robustheit gegen Angriffe zu evaluieren. Neue Kriterienkataloge zusammen mit effizient anwendbaren Zertifizierungsschemata und automatisierten Prüfverfahren sowohl für generative KI-Verfahren im engeren Sinn als auch für generative KI-Ökosysteme sind zu entwickeln. Aufgrund der hohen Dynamik der zu zertifizierenden KI-basierten Artefakte sind neue Vorgehensweisen erforderlich, um kontinuierlich und gleichzeitig effizient und nachvollziehbar die gewünschten Sicherheitsnachweise zu führen. Zudem werden neue Ansätze für formale Beweistechniken zu erforschen und zu erproben sein, um mit formal verifizierten Basiskomponenten zusammen mit formal verifizierten Prozessen zur Adaption und kontrollierbaren Weiterentwicklung von generativen KI-Ökosystemen eine Vertrauensbasis für die Beherrschbarkeit solcher Systeme zu etablieren.

Zur kontinuierlichen Zertifizierung und sicheren Weiterentwicklung generativer KI-Systeme werden automatisierte Test- und Prüfverfahren erforderlich. Penetrationstests müssen auf KI ausgeweitet (Adversarial Testing) werden, um Manipulationen aufzudecken. Gleichzeitig werden formale Methoden an Bedeutung gewinnen. Mit Theorem Proving (etwa Coq, Isabelle/HOL), Model Checking (TLA+) oder spezialisierten Tools wie Marabou können zum Beispiel bestimmte Eigenschaften neuronaler Netze mathematisch verifiziert werden. Ziel muss es sein, sicherheitskritische KI-Basiskomponenten formal zu zertifizieren und so eine „Chain of Trust“ bereitzustellen. Eine Herausforderung liegt vor allem in der Skalierbarkeit und Automatisierung dieser Verfahren sowie in der Integration in industrielle und behördliche Abläufe.

**Empfehlung 4: Neue Ansätze für das Fine-Tuning und den Betrieb von generativer KI**

Die am Markt verfügbaren mächtigen LLM sollten für die deutsche Wirtschaft und öffentliche Verwaltung effizient, schnell und vor allem sicher nutzbar gemacht werden. Nur dann kann die dringend erforderliche Transformation hin zu KI-unterstützter Digitalisierung mit der erforderlichen hohen Geschwindigkeit vorangetrieben werden. Dies kann gelingen, wenn die entsprechenden KI-Systeme in vertrauenswürdigen Umgebungen kontrolliert betrieben sowie kontrolliert und zugeschnitten (weiter)trainiert werden. Um Souveränität über Fine-Tuning, Betrieb und Adaption von LLM zu ermöglichen, sollte gezielt in die dafür erforderlichen Technologien und Lösungen investiert werden. In agilen Forschungskooperationen zwischen Wissenschaft und Industriepartnern müssen Lösungen entwickelt und in Reallaboren erprobt werden, die den Aufbau vertrauenswürdiger Umgebungen auf existierenden Cloud-Infrastrukturen ermöglichen. Ziel muss es sein, die auf diesen Umgebungen laufenden KI-Modelle und das gesamte KI-Ökosystem nachvollziehbar sicher und vertrauenswürdig zu betreiben – im Einklang mit Vorgaben wie der DSGVO.

**Empfehlung 5: Pilotprojekt für ein vertrauenswürdiges Verwaltungs-KI-Ökosystem**

Es wird empfohlen, in einem Pilotprojekt eine Blaupause für ein vertrauenswürdiges Verwaltungs-KI-Ökosystem und dessen nachweislich sicheren Betrieb auf Standard-Cloud-Infrastrukturen zu entwickeln. Es geht hierbei nicht um die Bereitstellung einer KI Plattform. Vielmehr soll eine erweiterbare KI-Betriebs-Infrastruktur entstehen, die es erlaubt, marktgängige Produkte sicher einzubinden, und gleichzeitig KI-Verfahren zusammen mit den erforderlichen Daten kontrolliert, nachvollziehbar und transparent so zu nutzen, dass die Einhaltung von regulatorischen Vorgaben nachgewiesen werden. Dazu könnten verfügbare Open-Source-KI-Modelle genutzt und mit domänenspezifischem, kuratiertem Wissen ausgestattet werden. Der Prototyp sollte Konzepte integrieren, die es ermöglichen, die Korrektheit der generierten Ausgaben sicher nachzuvollziehen. Durch den Einsatz moderner Sicherheitskonzepte soll nachprüfbar sichergestellt werden, dass keine sensiblen Daten in unautorisierte Hände gelangt. Die Verantwortlichen behalten die Kontrolle über den Zugriff auf bereichsspezifisches Wissen, aber zusätzlich wird ein sicherer, kontrollierter Informationsaustausch über Verantwortungsgrenzen hinweg ermöglicht. Es wird geraten, mittels eines agilen Ansatzes vorzugehen. Mit einem Minimum Viable Product (MVP) unter Verwendung öffentlich verfügbarer Daten könnte innerhalb weniger Wochen ein erstes, funktionales System bereitgestellt und evaluiert und danach kontrolliert schrittweise mit weiteren, geprüften Daten, zum Beispiel aus einem Verwaltungsbereich, trainiert und verfeinert werden. Der Prototyp sollte eng durch das BSI und Datenschutzbehörden begleitet werden, sodass die entstehende Plattformtechnologie für weitere Anwendungsfälle in Verwaltung und Unternehmen verfügbar gemacht werden kann.

Es wird weiterhin empfohlen, sich eng mit den Aktivitäten des Beratungszentrums für Künstliche Intelligenz (BeKI) des Bundesministerium des Innern und für Heimat (BMI) zum „Marktplatz der KI-Möglichkeiten“ abzustimmen. Dies ist die zentrale Matching-Plattform[6] für KI-Systeme in der Bundesverwaltung, um Bürgerinnen und Bürger, Wirtschaft,

Forschung sowie Ministerien und Behörden zu vernetzen. Zudem sollte eine enge Zusammenarbeit mit dem KI-Portal KIPITZ des ITZBund angestrebt werden, um auch dieses Portal mit den entsprechenden Sicherheitskonzepten auszustatten. Über KIPITZ können bereits heute behördeneigene Wissensdatenbanken mittels der RAG-Technik angebunden werden. Während KIPITZ derzeit für jede Behörde noch eine eigene Plattform zur Nutzung anbietet, könnte mit den zu entwickelnden Schutzkonzepten der Blaupause auch eine kontrollierte gemeinsame Nutzung von Daten über Behördengrenzen hinweg ermöglicht und damit ein deutlicher Mehrwert geschaffen werden. Weiterhin wird nahegelegt, die Lösungen auch in den zu etablierenden KI-Reallaboren zu erproben.

Übergreifend wird dringend **empfohlen, die bereits existierenden, häufig stark fragmentierten Maßnahmen stärker zu koordinieren und den Austausch zu fördern.** Über vertrauenswürdige KI-Ökosystem-Plattformen kann neben einem generellen Erfahrungsaustausch auch die technisch abgesicherte, kontrollierte und ressort- oder abteilungsübergreifende Nutzung trainierter KI-Modellen und Anwendungen angeboten werden.

## Fazit

Das Papier zeigt neben den neuen, teilweise noch kaum erforschten Risiken, die durch generative KI entstehen, auch insbesondere die Chancen auf, die mit dem Einsatz von generativer KI sowohl für die Erhöhung der Cybersicherheit als auch für eine schnelle, abgesicherte digitale Transformation in Behörden und Unternehmen verbunden sind. Klar ist jedoch auch, dass viele, aus Sicherheitssicht sehr hilfreiche Anwendungsbereiche der generativen KI dieses Potenzial nur dann entfalten können, wenn KI vertrauenswürdig arbeitet und für deren Nutzung vertrauenswürdige Trainings- und Betriebsumgebungen etabliert werden. Ansonsten verwandeln sich vermeintliche Chancen in zusätzliche Risiken.

Politisches Handeln ist dringend erforderlich, um die Position Deutschlands in Bezug auf wertvolles Domänenwissen und die entsprechenden wertvollen Domänendaten zu nutzen, um dediziert trainierte, vertrauenswürdige generative KI-Ökosysteme zu etablieren und damit die Abhängigkeiten von den großen Technologieanbietern zu verringern. Mit der Umsetzung der Empfehlungen könnte die Politik die Rahmenbedingungen für den schnelleren und effektiveren Einsatz von KI insbesondere auch für KMU verbessern. Sie könnte Initiativen stärken bzw. initiieren, sodass die Chancen genutzt und Risiken minimiert werden.

Vertrauenswürdige generative KI-Ökosysteme bieten die Chance, mit automatisiert einsetzbaren Techniken die Geschäftsmodelle von Angreifeden wirksam und effizient zu zerstören oder zumindest substanziell zu beeinträchtigen.

## Empfehlungen zur Umsetzung von Ambitionen des Koalitionsvertrags der neuen Bundesregierung

Das vorliegende Positionspapier liefert Empfehlungen zur Umsetzung konkreter Maßnahmen für viele der Ambitionen, die sich die neue Bundesregierung im Koalitionsvertrag vorgenommen hat. Zentrale Themen des Koalitionsvertrags betreffen die **Stärkung der Cybersicherheit** im Allgemeinen (Zeile 2185ff, 2547ff, 2675ff ) und des Mittelstandes im Besonderen (Zeile 284), die Stärkung der **digitalen Souveränität** (Zeile 2140 ff) und die Modernisierung und **Digitalisierung der Verwaltung** (Zeile 1794 ff, 1857ff). Die Ambition der **Stärkung der Zukunftstechnologie Künstliche Intelligenz** findet ihren Niederschlag an vielen Stellen im Vertrag (u. a. Zeilen 88, 108, 167, 584, 1857, 2161, 2167, 2263, 2509, 2851, 3500). Deutschland soll eine **KI-Nation** werden (Zeile 88, 108), KI soll umfassend genutzt werden (Zeile 2167) und es sollen **KI-Sprunginnovationen** entstehen (Zeile 2263). Dabei sollen ein **offenes Datennutzungsverständnis** gefördert, berechtigte Interessen bewahrt (Zeile 2246) und **KI-Ökosysteme** auch in sehr sensitiven Bereichen zugänglich gemacht werden, wie für Sicherheitsbehörden (Zeile 2851), dem Gesundheitsbereich (Zeile 3500), Bildung/Schule (Zeile 2335) oder auch in der Verwaltung (Zeile 2161). **Das Positionspapier umfasst fünf Empfehlungen zur Gestaltung einer digital souveränen KI-Nation basierend auf vertrauenswürdigen KI-Ökosystemen, um diese Ambitionen gezielt anzugehen.** Dem BSI wird dabei eine wichtige Rolle zukommen. Die empfohlenen Maßnahmen zur Umsetzung adressieren Aufgaben, die in unterschiedlichen Ressorts federführend betrieben werden sollten, neben BMI und dem neu zugeschnittenen Ministerium für Forschung, Technologie und Raumfahrt (Zeile 4575) ist dies insbesondere auch das neue Ministerium für Digitales und Staatsmodernisierung (Zeile 4564).

Die **Empfehlung 1** – Erforschen neuer Ansätze für generative KI-Ökosysteme – adressiert zum einen die Absicht, KI Sprunginnovationen (Zeile 2263ff) zu unterstützen und zu nutzen und zum anderen das generelle Ziel, die digitale Souveränität zu stärken und die Robustheit von KI-basierten Ökosystemen bzw. Infrastrukturen zu erhöhen, indem Abhängigkeiten reduziert werden. Die Empfehlung adressiert auch das Ziel, zu einem offeneren Umgang mit Daten zu kommen, aber gleichzeitig Datenschutzinteressen zu wahren (Zeile 1858) und Privacy Enhancing Technologies (PET) zum Einsatz zu bringen. Sie unterstützt gleichzeitig die Position der BfDI, neue Ansätze für einen ermöglichenden Datenschutz zu verfolgen.

Die **Empfehlung 2** – Entwicklung vertrauenswürdiger, generativer Cybersicherheitslösungen – adressiert zum einen die Stärkung der digitalen Souveränität im Bereich vertrauenswürdiger KI-Systeme und zum anderen konkrete Empfehlungen zur Stärkung der Cybersicherheit im Mittelstand (Zeile 284) durch die Entwicklung von KI-Assistenzsystemen für die Domäne der Cybersicherheit. Die Empfehlung, Werkzeuge zur automatisierten Unterstützung von Compliance-Prozessen zu entwickeln, adressiert direkt das Ziel (Zeile 285), Unternehmen bei der Umsetzung des CRA zu unterstützen. Mit den zu entwickelnden KI-Lösungen können zudem vertrauenswürdige Lösungen für Sicherheitsbehörden bereitgestellt werden (Zeile 2636).

Die **Empfehlung 3** – Neue Testverfahren und Zertifizierungsschemata – ist ein zentraler Baustein, um den Anspruch zu erfüllen, die Resilienz zu stärken (Zeile 2185) sowie integrierte KI-Lösungen in Bezug auf ihre Vertrauenswürdigkeit zu prüfen (Zeile 2160). Mit den empfohlenen Maßnahmen wird zudem der Aufbau der Deutschen Verwaltungscloud (Zeile 2164) adressiert.

Die **Empfehlung 4** – Neue Ansätze für den Betrieb von generativer KI – adressiert den Anspruch, über Hub-Strukturen (Zeile 2499) Innovationsräume zu schaffen. Durch den Aufbau und Betrieb von vertrauenswürdigen KI-Ökosystemen in Hub-Strukturen und KI-Reallaboren (Zeile 2266) wird der Transfer zwischen Forschung und Wirtschaft, sowie Start-up-Szene und öffentlicher Verwaltung gefördert, und die souveräne, vertrauenswürdige Nutzung von KI beschleunigt. Die Umsetzung der Empfehlung unterstützt auch das Ziel (Zeile 1858), Daten zur strategischen Steuerung, Modellierung und Wirkungskontrolle besser zu bündeln und besser zu nutzen.

Die **Empfehlung 5** – Pilotprojekt für ein vertrauenswürdiges Verwaltungs-KI-Ökosystem – adressiert das Ziel der Modernisierung und Digitalisierung der Verwaltung (Zeile 1794 ff), Verwaltungsprozesse zu automatisieren (Zeile 1862 ff), zu beschleunigen und effizienter zu gestalten – insbesondere mit Künstlicher Intelligenz – und den Zugang zu und die Verknüpfung von relevanten Daten sicherzustellen. Das Pilotvorhaben liefert sowohl einen Beitrag zur Verwaltungsmodernisierung als auch durch eine konsequente Open-Source Ausrichtung einen Umsetzungsbeitrag zu einer Open-Source Strategie, die durch die im Koalitionsvertrag benannten Institutionen (Zeile 2172) zu entwickeln sein wird.

1 McKinsey & Company. (2023, Juni 14). Das wirtschaftliche Potenzial der generativen KI: Die nächste Produktivitätsgrenze [Bericht]. Abgerufen von https://www.mckinsey.com
2 Marchal, S., & Nawrotek, B. (2024). Anwendung von künstlicher Intelligenz in der Cybersicherheit. (Traficom Research Reports No. 07/2024). Finnische Agentur für Verkehr und Kommunikation Traficom. https://www.kyberturvallisuuskeskus.fi/en/publications/applying-artificial-intelligence-cybersecurity
3 Unter einem KI-Ökosystem verstehen wir die Gesamtheit der für das Training und den operativen Betrieb erforderlichen Komponenten, wie generative Modelle, Trainingsverfahren, Trainingsdaten, APIs, Ausführungsumgebungen oder auch Infrastrukturkomponenten wie Cloud-Plattformen oder auch Endgeräte.
4 Dawson, A. (Release Lead), & Wilson, S. (2023). OWASP Top 10 für LLM-Anwendungen (Version 1.1). OWASP Foundation. https://LLMtop10.com
5 (1) Unterstützung proaktives Handeln, um Angriffs-Resilienz zu erhöhen, (2) Assistenz für operative Routineaufgaben für Sicherheitsverantwortliche, (3) Verbesserung der Sicherheitsqualität von Software.
6 https://maki.beki.bund.de/a/bmi-makimo-app?kiosk

**Wissenschaftliche Arbeitsgruppe Nationaler Cyber-Sicherheitsrat**

Die Wissenschaftliche Arbeitsgruppe wurde im Oktober 2018 gegründet und ist Mitglied des Nationalen Cyber-Sicherheitsrats. Sie berät aus Perspektive der Forschung zu Entwicklungen und Herausforderungen im Hinblick auf eine sichere, vertrauenswürdige und nachhaltige Digitalisierung.

Mitglieder der Wissenschaftlichen Arbeitsgruppe sind: Thomas Caspers, Prof. Dr. Gabi Dreo Rodosek, Prof. Dr. Claudia Eckert, Prof. Dr. Jörn Müller-Quade, Prof. Dr.-Ing. Christof Paar, Prof. Dr. Alexander Roßnagel, Prof. Dr. Michael Waidner

Hauptautorin des Impulspapiers „Generative Künstliche Intelligenz und ihre Auswirkungen auf die Cybersicherheit“: Prof. Dr. Claudia Eckert