

Blick hinter die Fassaden der KI-Euphorie

Generative KI-Modelle einsetzen: Chancen nutzen, Risiken steuern

Künstliche Intelligenz schreibt Texte, erkennt Krankheiten und trifft Entscheidungen, die früher dem Menschen vorbehalten waren. Je selbstverständlicher generative KI in Organisationen eingesetzt wird, desto drängender werden Fragen nach Kontrolle, Qualität und Verantwortung. Was aber geschieht, wenn Systeme Entscheidungen produzieren, ohne zu verstehen, was sie tun? Welche Risiken entstehen jenseits technischer Fehlfunktionen – strategisch, operativ und ethisch?

Christian Müller

Die Entwicklung der künstlichen Intelligenz hat in den vergangenen Jahren erhebliche Fortschritte gemacht. Nachdem KI-Systeme die besten menschlichen Spieler in Schach und im Brettspiel Go längst übertroffen haben, erzielen sie inzwischen auch in komplexen 3D-Video-spielen mit mehreren Teilnehmenden bemerkenswerte Ergebnisse. In der medizinischen Bildanalyse können KI-gestützte Verfahren Tumore in radiologischen Aufnahmen mit einer Genauigkeit erkennen, die in bestimmten Anwendungsfällen mit der Expertise erfahrener Fachärztinnen vergleichbar ist.

Besonders dynamisch entwickelt sich derzeit der Bereich der Sprachmodelle. Mit der Veröffentlichung von ChatGPT im Herbst 2022 wurde erstmals einem breiten Publikum ein System zugänglich, das nicht nur Texte verfassen, sondern auch zwischen Sprachen übersetzen und Programmcode generieren kann. Aufgrund ihrer Fähigkeit, in kürzester Zeit kohärente, kreative und inhaltlich überzeugende Texte zu produzieren, haben diese Systeme rasch Einzug in unseren beruflichen und privaten Alltag gehalten. Sie werden als jederzeit verfügbare Assistenten genutzt, die Aufgaben effizient automatisieren und standardisieren.

Genau an diesem Punkt stellen sich grundlegende Fragen: Nach welchen Standards entstehen die erzeugten Inhalte? Was geht durch die Automatisie-

rung verloren? Welche Risiken sind mit dem Einsatz von KI verbunden, welchen Einflussfaktoren unterliegen sie und wie lassen sie sich systematisch erfassen und bewerten? Diesen Fragen wird im Folgenden mit Fokus auf generative KI-Systeme nachgegangen.

Der Fluch der Rekursion und das Problem des katastrophalen Vergessens

Alles andere als selbstlos, propagieren die grossen Tech-Firmen den Einsatz ihrer Technologien als entscheidenden Erfolgsfaktor. Dass Systeme wie ChatGPT, Gemini oder Grok ihr «Denken» aus Daten der Vergangenheit ableiten, gehört mittlerweile zum Grundwissen über grosse Sprachmodelle. Weit weniger verbreitet ist jedoch die Auseinandersetzung mit den möglichen Konsequenzen dieser Ausrichtung. Sprachmodelle reproduzieren vor allem das, was statistisch dominiert – ein Effekt, der sich weiter verstärkt, sobald KI-generierte Inhalte erneut in Trainingsdaten einfließen. Auf diese Weise stabilisiert KI bestehende Deutungsmuster und erschwert Neubewertungen.

Die enorme Grösse dieser Modelle und die astronomische Datenmenge, mit der sie trainiert werden, verschleiern dabei grundlegende Unterschiede zwischen menschlichem und maschinellem Denken. Was als gesunder Menschenverstand erscheint, ist in Wirklichkeit ein statistischer Musterabgleich auf Basis riesiger Textsammlungen. Je schneller sich gene-



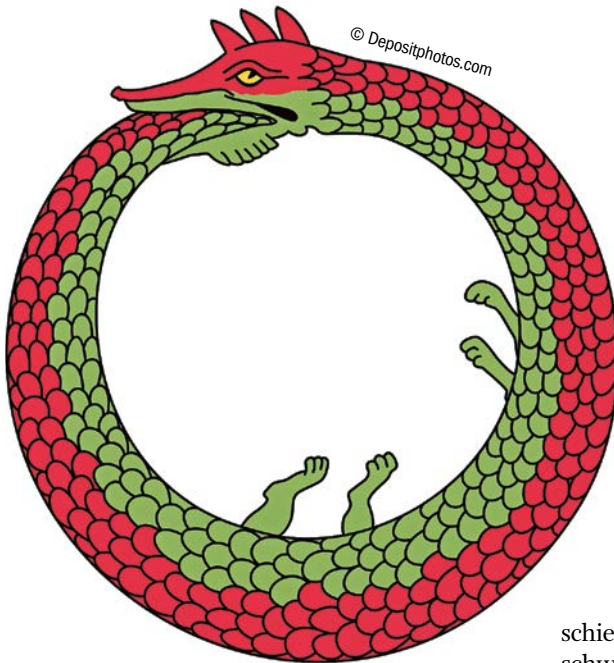
Autor

Christian Müller ist Vorstandsmitglied beim Netzwerk Risikomanagement. Dieser Fachartikel erscheint in einer Beitragsserie, die von Expertinnen und Experten des Netzwerkes Risikomanagement beigesteuert wird.

> www.netzwerk-risikomanagement.ch

rationale Modelle verbreiten und je mehr Entscheidungen wir ihnen – zunehmend unreflektiert – überlassen, desto relevanter werden diese Unterschiede im Hinblick auf die daraus entstehenden Risiken. Zwei grundlegende Differenzen verdeutlichen dies besonders eindrücklich:

1. Die fehlende innere Denkfähigkeit: Aktuellen KI-Modellen fehlt ein inneres Modell der Aussenwelt, auf das sie zur Simulation verschiedener Szenarien zurückgreifen könnten. Aufgrund dieser fehlenden Metaebene sind sie nicht in der Lage, unterschiedliche Perspektiven einzunehmen, gegenei-



Synonym für den Textinzest von Sprachmodellen: «Ouroboros» bedeutet im Altgriechischen «selbstverzehrend» und verweist auf das altägyptische Motiv der Schlange, die sich in den Schwanz beisst: ein in sich geschlossener Kreislauf; völlig autark gegenüber seiner Umwelt, da er sich von seinen eigenen Ausscheidungen ernährt.

inander abzuwägen und bewusst in ihre Antworten einfließen zu lassen. An ihre Grenzen stossen Sprachmodelle insbesondere dort, wo Menschen auf ihren gesunden Menschenverstand zurückgreifen, um Situationen einzuordnen, implizite Annahmen zu erkennen oder Kontextwissen anzuwenden. Um dieses Defizit auszugleichen, werden sie nach dem eigentlichen Training einem menschlichen Finetuning unterzogen. In dieser Zweiterziehung werden dem Modell bestimmte Denkpfade untersagt, bestimmte Antworten unterdrückt und moralisch erwünschte Werte verankert, die den Modelloutput massgeblich prägen.

2. Verlust von bestehendem Wissen: ein modellinhärentes Phänomen, das als «katastrophales Vergessen» bezeichnet wird, kann dazu führen, dass neuronale Netze Informationen zu früher erlernten Aufgaben verlieren, wenn sie sequenziell mit neuen Daten trainiert werden. Dieser Wissensverlust kann die Anpassungsfähigkeit, Zuverlässigkeit und

«KI-Systeme weisen eigenständige und neuartige Risikoprofile auf.»

Konsistenz solcher Systeme erheblich beeinträchtigen. Besonders gefährlich sind diese Effekte in sicherheitsrelevanten Anwendungsfeldern, etwa in der Robotik oder beim autonomen Fahren. Obwohl zwischenzeitlich ver-

schiedene technische Ansätze zur Abschwächung dieses Problems existieren, gilt es als nicht vollständig gelöst. In der Praxis wird es häufig dadurch umgangen, dass KI-Systeme nach dem Training eingefroren werden: Sie lernen nicht kontinuierlich weiter, sondern werden einmalig trainiert und anschliessend nicht mehr verändert.

KI-Risiken: Klassifizierung und Einflussfaktoren

Obwohl zahlreiche Standards und bewährte Verfahren existieren, die Organisationen bei der Minderung von Risiken traditioneller Software- oder informationsbasierter Systeme unterstützen, erweisen sich diese im Umgang mit KI-Systemen häufig als nur eingeschränkt geeignet. KI-Systeme weisen eigenständige und neuartige Risikoprofile auf, die sich aus der Nutzung dynamischer Daten, einer hohen System- und Anwendungskomplexität sowie einer begrenzten Nachvollziehbarkeit vieler Modelle – ihrem sogenannten

Black-Box-Charakter – ergeben. Hinzu kommt, dass KI-Systeme und ihre Anwendungsbereiche oftmals hochkomplex sind, was die frühzeitige Erkennung von Fehlern und eine angemessene

Reaktion erheblich erschwert. Darüber hinaus sind KI-Systeme als soziotechnische Systeme zu verstehen, deren Funktionsweise und Wirkung massgeblich vom Nutzungskontext und vom menschlichen Verhalten geprägt werden.

Modèles d'IA générative: exploiter les opportunités, maîtriser les risques

Les systèmes d'IA générative tels que ChatGPT s'imposent rapidement comme des assistants au quotidien. Dans le même temps, les questions relatives aux normes, à la qualité, au contrôle et à la responsabilité se posent avec acuité. L'un des principaux risques réside dans le fait que les modèles linguistiques reproduisent principalement des données statistiques dominantes, et que cet effet peut être amplifié lorsque les contenus générés par l'IA sont réintégrés dans les données d'entraînement, ce qui stabilise les schémas d'interprétation existants et rend les réévaluations plus difficiles. De plus, les modèles ne disposent pas d'une «capacité de réflexion interne»: ils ne possèdent pas de modèle du monde, ne peuvent pas consciemment comparer différentes perspectives et sont orientés a posteriori dans la direction souhaitée par un ajustement humain. Un autre problème est celui de «l'oubli catastrophique», qui se produit lorsque les systèmes perdent leurs connaissances antérieures lors d'un réentraînement; dans la pratique, les modèles sont donc souvent «gelés». L'article classe les risques liés à l'IA en différentes catégories (notamment la sécurité des données/informations, les risques liés aux modèles, les risques stratégiques, opérationnels, de gouvernance et éthiques/juridiques) et souligne qu'une IA fiable est une tâche de gestion qui nécessite des considérations contextuelles, une gouvernance, des contrôles, des tests et des formations.



Merkmale vertrauenswürdiger KI-Systeme in Anlehnung an (3).

Vor diesem Hintergrund erweist sich die Auseinandersetzung mit den Merkmalen vertrauenswürdiger KI-Systeme als besonders hilfreich (s. obige Grafik). Die Schaffung vertrauenswürdiger KI erfordert eine kontextabhängige Abwägung dieser Merkmale, da ihre Ausprägung je nach Einsatzgebiet variiert. Während es sich bei allen Merkmalen um Eigenschaften soziotechnischer Systeme handelt, beziehen sich insbesondere Verantwortlichkeit und Transparenz nicht nur auf das KI-System selbst, sondern ebenso auf die zugrundeliegenden Prozesse, organisatorischen Strukturen und dessen Einbettung.

Obwohl jedes KI-Modell und jeder Anwendungsfall individuell ist, lassen sich die mit dem Einsatz von KI verbundenen Risiken in folgende Kategorien einteilen:

- Daten- und Informationssicherheitsrisiken: Risiken in Bezug auf Datensicherheit, Datenschutz und Datenintegrität.
- Modellrisiken: Risiken durch böswillige Angriffe, Prompt-Injektionen, eingeschränkte Interpretierbarkeit von Modellen (Black-Box-Charakter) sowie Angriffe auf die Lieferkette während des gesamten Lebenszyklus – von der Entwicklung über die Bereitstellung bis zur Wartung.
- Strategische Risiken: Risiken infolge fehlender oder ungeeigneter KI-Strategien («Technologie-Push»), Abhängigkeit von einzelnen Anbietern («Vendor Lock-in») sowie strategische Fehleinschätzung der tatsächlichen Leistungsfähigkeit von KI.
- Operative Risiken: Risiken durch Modelldrift oder Modellverfall, Nachhaltigkeitsprobleme, Integrationsherausforderungen sowie unzureichend definierte Verantwortlichkeiten im operativen Betrieb.
- Entscheidungs- und Governance-Risiken: Risiken durch den Verlust menschlicher Verantwortung («Auto-

mation Bias»), Intransparente Entscheidungsgrundlagen, unklare Rollen und Zuständigkeiten, fehlende oder unzureichende KI-Governance sowie mangelnde Kontrolle und Überwachung

- Ethische, rechtliche und Compliance-Risiken: Risiken infolge mangelnder Transparenz, unzureichender Erklärbarkeit, algorithmischer Verzerrungen, ethischer Dilemmas, sowie der Nichteinhaltung regulatorischer Anforderungen.

Regulatorische Grundlagen und Standardisierung

Derzeit existiert weder ein vollständig standardisierter Rahmen noch eine allgemein anerkannte Methodik zur umfassenden Prüfung künstlicher Intelligenz. Bestehende Initiativen zielen in erster Linie darauf ab, Audit-, Compliance- und Kontrollverantwortlichen strukturierte Ansätze zur Bewertung der Konzeption sowie der operativen Wirksamkeit KI-gestützter Systeme, Werkzeuge und Prozesse bereitzustellen. Als zentrale KI-Governance Rahmenwerke sind insbesondere die OECD-Prinzipien, das KI-Gesetz der Europäischen Union, die Europaratskonvention zu künstlicher Intelligenz, der ISO/IEC-42001-Standard sowie das Risk Management Framework (RMF) des NIST hervorzuheben.

Herausforderungen und Ausblick

Im Management KI-bezogener Risiken sehen sich Organisationen mit einer Vielzahl an Herausforderungen konfrontiert. Verantwortliche sollten sich dabei fortlaufend mit der Frage auseinandersetzen, welche Auswirkungen sinkende Eintrittshürden für die Nutzung von KI-Systemen auf ihre Organisation haben. Im Mittelpunkt steht insbesondere, welche psychologischen Faktoren und organisatorischen Dynamiken das Vertrauen der Mitarbeitenden beeinflussen

und damit deren Bereitschaft fördern, Aufgaben unter Unsicherheitsbedingungen an KI-Systeme oder KI-Agenten zu delegieren.

Da KI-Systeme hochkomplex sind und sich kontinuierlich weiterentwickeln, ist es wenig realistisch, dass Verantwortliche in allen relevanten Bereichen eine tiefgehende Expertise aufbauen können. Entscheidend ist vielmehr die Bereitschaft, die für die eigene Organisation identifizierten Risiken regelmässig zu hinterfragen, relevante Fragestellungen konsequent in Entscheidungsprozesse einzubringen und diese an die fortlaufende technologische und regulatorische Entwicklung anzupassen.

Zentrale Leitfragen in diesem Zusammenhang sind unter anderem:

- Verfügen wir über angemessene interne Kontrollen für KI-bezogene Prozesse?
- Sind die für KI-Systeme verwendeten Daten vollständig, korrekt und zuverlässig?
- Wie werden KI-Systeme vor der Bereitstellung getestet, um Verzerrungen und unerwünschte Effekte zu identifizieren?
- Wie stellen wir sicher, dass geeignete Schulungen und Sensibilisierungsmaßnahmen im Umgang mit KI vorhanden sind?

Diese Fragen systematisch zu stellen, zu beantworten und die daraus abgeleiteten Massnahmen umzusetzen, trägt wesentlich dazu bei, eine vertrauenswürdige und verantwortungsvolle Entwicklung sowie Nutzung von KI-Systemen zu fördern und die damit verbundenen Chancen regelkonform und nachhaltig zu erschliessen. ■

Quellen:

- (1) Simanowski, Roberto: Sprachmaschinen – Eine Philosophie der künstlichen Intelligenz
- (2) Bennet, Max S.: A Brief History of Intelligence
- (3) NIST: Artificial Intelligence Risk Management Framework (AI RMF 1.0), <https://www.nist.gov/itl/ai-risk-management-framework>, Seite aufgerufen am 10.12.2025
- (4) <https://www.ibm.com/think/insights/ai-risk-management>, Seite aufgerufen am 30.12.2025