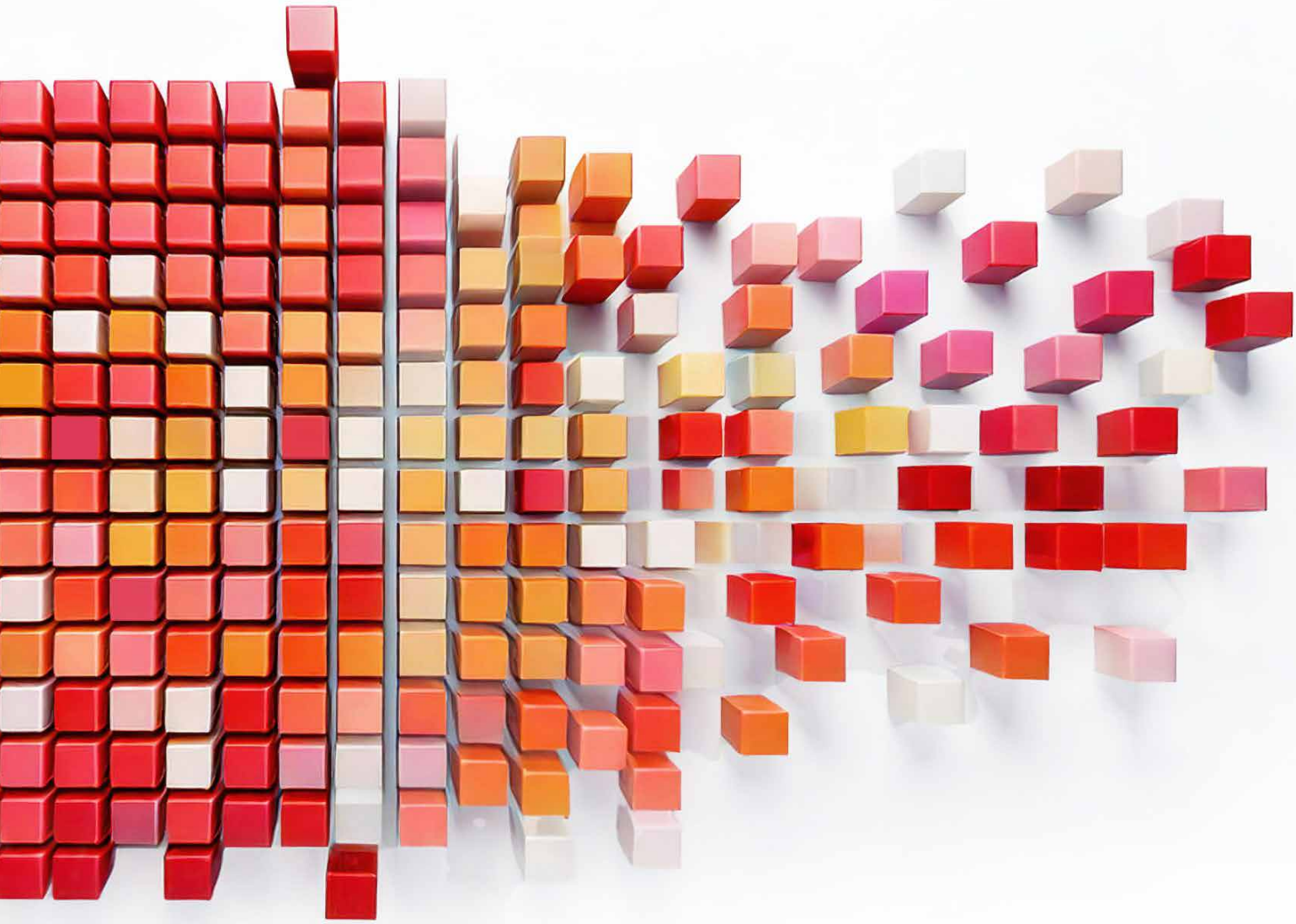
www.pwc.de

Agentic AI – alles unter Kontrolle?

Herausforderungen in Zeiten autonom agierender KI-Bots



pwc

Management Summary

- Agentic AI ist die nächste Evolutionsstufe für die Anwendung von KI. Die geschickte Verkettung bestehender KI-Modelle und die Integration mit externen Systemen eröffnet Unternehmen neue Anwendungspotenziale.
- Wesentliche Merkmale von Agentic AI sind ein höherer Grad an Autonomie und die Möglichkeit, eigenständig Entscheidungen zu treffen und Aktionen auszulösen.
- Zusammen mit den Effizienzvorteilen durch Agentic AI steigen auch die Risiken. Da menschliche Aufsicht reduziert wird, müssen Kontrollen stärker in die Ausgestaltung der KI-Systeme und das Agentic AI Design rücken. Nicht oder zu spät erkannte Fehler haben drastischere Auswirkungen.
- Unternehmen müssen ihre KI-Governance erweitern und die Risiken managen, um von Agentic AI profitieren zu können.
- Eine wesentliche Herausforderung ist die richtige Balance zwischen autonomer KI und menschlicher Kontrolle.
- Regelwerke wie der AI Act gehen (noch) nicht explizit auf den „Agentic-Trend“ ein. Der AI Act greift aber in jedem Fall bei Hochrisikosystemen.

Einleitung

„I'm sorry, Dave. I'm afraid I can't do that" – welche katastrophalen Auswirkungen der Kontrollverlust über eine künstliche Intelligenz haben kann, zeigte bereits 1968 der Filmklassiker 2001: Odyssee im Weltraum von Stanley Kubrick. Heute, im Jahr 2025, sind wir noch lange nicht bei einer KI Superintelligenz angekommen und ob es sie jemals geben wird, ist fraglich. Ein höherer Autonomiegrad von KI und eine komplexere Task-Bewältigung kommen aber in schnellen Schritten. Die Rede ist von Agentic AI – also KI-Systemen, die (zumindest theoretisch) weitgehend autonom agieren, eigenständig Aufgaben lösen und sich adaptiv weiterentwickeln können. In der Praxis sind solche Systeme bislang meist als Multiagentenlösungen implementiert und für konkrete Zwecke ausgelegt.

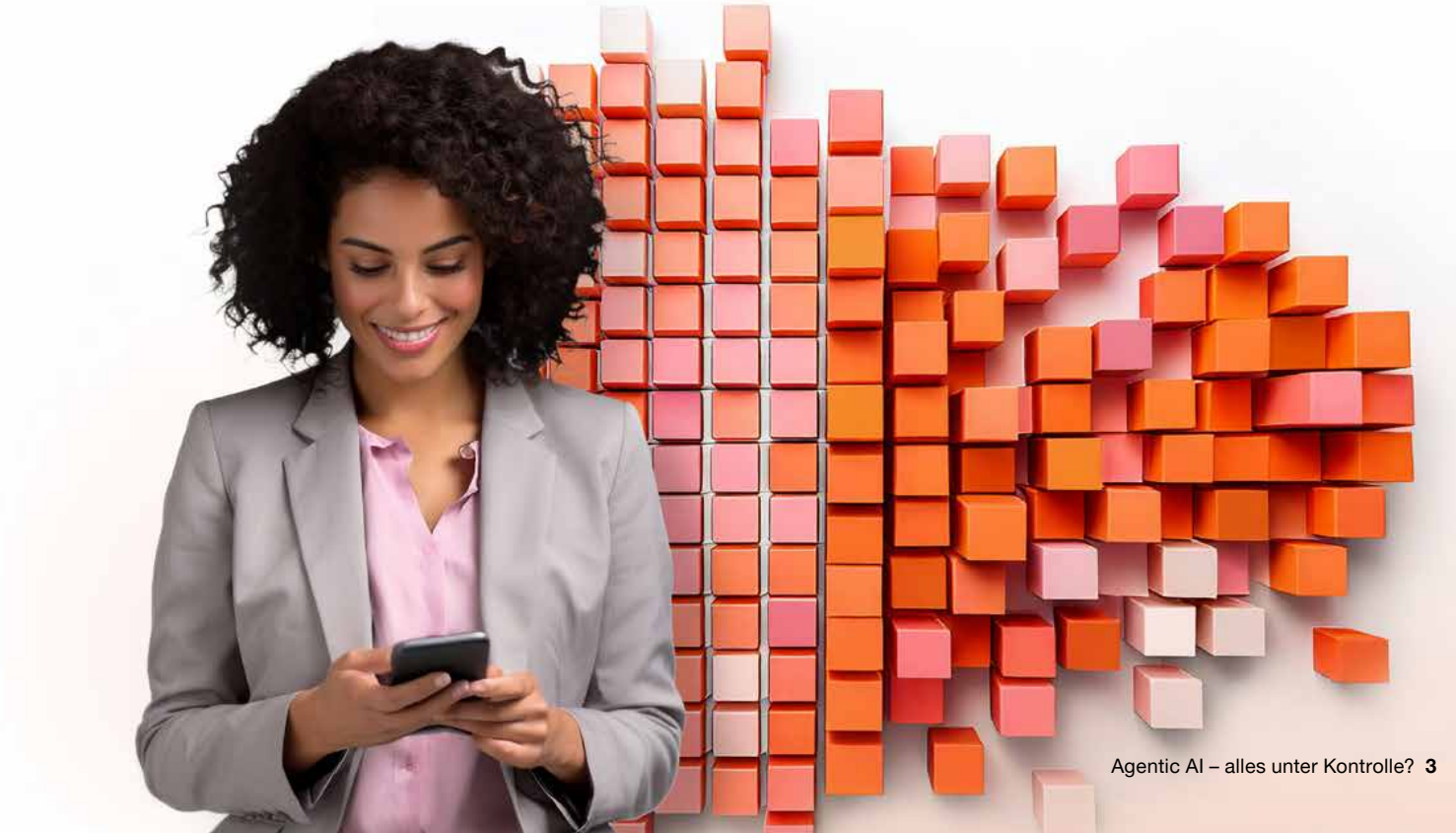
Das neue KI-Paradigma verstärkt einen Autonomietrend, der schon in vielen Bereichen Einzug gehalten hat – sei es in Form von Robotik-getriebenen Produktionssystemen, Systemen zur Verkehrssteuerung, Lieferdrohnen oder autonomen Erntemaschinen in der Landwirtschaft.

Die zugrundeliegende Technologie hinter Agentic AI ist nicht neu, bestehende KI-Modelle wie LLMs spielen weiterhin die Hauptrolle. Die Art und Weise, wie die bestehenden Bausteine verkettet, ineinander geschaltet und mit weiteren Systemen verknüpft werden, birgt aber eine neue Qualität. Sie ermöglicht das Lösen komplexer, mehrstufiger Probleme – vorausgesetzt, die KI bekommt zu einem gewissen Grad freie Hand. Für Unternehmen aller Branchen eröffnen sich dadurch vielfältige Anwendungsmöglichkeiten, sei es im Kundenservice, in der Logistik, oder in den Bereichen Sales und Marketing.

Agentic AI-Anwendungen sind in der Lage, bestimmte Entscheidungen eigenständig durchzuführen und weiterführende Aktionen auszulösen. Sie setzen deshalb einen höheren Grad an Autonomie voraus. Wie lässt sich aber die richtige Balance finden zwischen einer KI, die autonom agiert, und einer menschlichen Aufsicht, die angemessen kontrolliert? Was müssen Kontrollsysteme künftig leisten, wenn KI-Anwendungen autark bestimmte Aufgaben abarbeiten? Und was passiert in Fällen, in denen KI-getriggerte Prozesse gestoppt werden müssen?

Um die nächste Evolutionsstufe von KI-Systemen sicher und verantwortungsbewusst zu betreiben, müssen Unternehmen auch ihre AI Governance weiterentwickeln. Der Nutzen von Agentic AI erschließt sich erst, wenn die Risiken kontrollierbar werden – und die KI eben nicht wie HAL 9000 außer Kontrolle gerät, den Dienst verweigert und massive Schäden anrichtet.

> Je mehr Verantwortung an KI-Systeme delegiert wird, desto engmaschiger müssen sie kontrolliert werden. Der Trend in Richtung Agentic AI erhöht die Relevanz einer robusten AI Governance.
>
> **Hendrik Reese**, Partner bei PwC Deutschland



Agentic AI – die nächste Evolutionsstufe von GenAI

Generative KI-Systeme haben in kurzer Zeit rasante Fortschritte gemacht. Mittlerweile ist eine Vielzahl von Modellen mit individuellen Stärken und Schwächen auf dem Markt. Darüber hinaus schreitet die Integration von GenAI-Systemen in unternehmerische Wertschöpfung und Prozesse voran. Da die KI-Modelle meist in der Cloud betrieben werden und über entsprechende Schnittstellen verfügen, ist die Anbindung an weitere IT-Systeme sehr einfach. Diese Ausgangsbedingungen haben sehr organisch die nächste Evolutionsstufe der KI eingeläutet: Agentic AI.

Wo verläuft die Grenze zwischen GenAI und Agentic AI?

Agentic AI zielt darauf ab, komplexe mehrstufige Probleme zu lösen. Dafür kommen potenziell mehrere Modelle zum Einsatz, die als KI-Bots bestimmte Teilaufgaben verantworten und mit externen Systemen interagieren. Die Interaktion mit Menschen hingegen, die bei Prompt-basierten GenAI-Systemen häufig im Mittelpunkt steht, wird dadurch reduziert. So versteht OpenAI unter Agentic AI zum Beispiel „AI Systems that can pursue complex goals with limited direct supervision“[1].

Wie unterscheidet sich Agentic AI von GenAI?

Klassische GenAI nimmt meist die Form von Assistenten an, die ad hoc Unterstützung bieten – sei es beim Formulieren einer E-Mail oder der Extraktion von Informationen aus großen Dokumenten. Die Interaktion läuft über einen Dialog oder zugrundeliegende Prompts. Agentic AI hingegen nutzt GenAI-Modelle, um autonome Bots zu kreieren. Sie gleichen vom Prinzip eher einem Staubsauger-Roboter, der einmal eingerichtet wird und dann kontinuierlich seine Aufgaben erfüllt. Agentic AI Bots treffen eigenständig Entscheidungen und können Handlungen in der Welt ausführen – abhängig davon, mit welchen Systemen sie gekoppelt sind und über welche Zugriffsrechte sie verfügen.

Aus technischer Sicht sind die Bausteine, auf denen Agentic AI aufsetzt, nicht neu. Die Innovation liegt vielmehr darin, wie bestehende GenAI-Modelle eingesetzt und verkettet werden. Anwendbar ist das Prinzip autonom agierender KI-Bots auf etliche Bereiche, es verspricht eine bessere Unterstützung für Nutzer in diversen Lebenslagen. Indem die Bots nicht nur auf Anfragen reagieren, sondern proaktiv bestimmte Aufgaben erfüllen, können sie uns potenziell eine Menge Arbeit abnehmen. Gerade für Unternehmen verspricht Agentic AI einen Produktivitätsboost und bietet die Chance, aus vorhandenen GenAI-Systemen noch mehr rauszuholen.

Beispiele für den Einsatz von Agentic AI

Kundenservice

AI Bots können beispielsweise für Online-Händler kontinuierlich die Kundenzufriedenheit im Blick behalten. Bei einer verspäteten Lieferung benachrichtigen sie die Kunden und bieten proaktiv einen Rabatt für die nächste Bestellung an.

Cyber Security

Agentic AI bietet neues Potenzial für die automatisierte Abwehr von Cyberangriffen. Agenten können Bedrohungen erkennen und sie abwehren. Auch die Angreifer können das Prinzip allerdings nutzen – z. B., um Sicherheitslücken zu identifizieren.

Healthcare

In Pflegesituationen wie dem betreuten Wohnen können AI Bots organisatorische Aufgaben übernehmen. Sie strukturieren den Tag der Bewohner, entwerfen Essens- und Transportpläne und erinnern Patient:innen an die Medikamenteneinnahme.

Neben individuell zugeschnittenen Agentic AI-Systemen für bestimmte Unternehmenszwecke stehen auch immer mehr universell einsetzbare Lösungen bereit – sei es Amazon Q, Microsoft Magentic One oder Agentforce von Salesforce. Die derzeit häufigsten Anwendungsfälle liegen demnach in den Bereichen Softwareentwicklung und genereller Computer-Nutzung. Es sind jedoch auch schon beispielsweise Systeme vertreten, die explizit für Robotik-Anwendungen ausgelegt sind.

[1] https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf.

Wie funktioniert Agentic AI?

Während bei GenAI die Ergebnisse meist im Dialog zwischen Menschen und KI-Modellen entstehen, spielt bei Agentic AI die initiale Zielsetzung eine größere Rolle. KI-Agenten bekommen ein übergeordnetes Ziel und verfolgen dann eigenständig die Schritte, um es zu erreichen. In der Praxis kommen dafür häufig manuell zusammengebaute Multiagentensysteme zum Einsatz. Das Problem wird in mehrere Schritte zerlegt, die dann jeweils von einem LLM bearbeitet werden. Ein gängiges Basismodell sieht folgende Schritte vor:

- **Perceive:** Das KI-System ruft Daten ab, die für die Erreichung des Ziels relevant sind. Diese Aufgabe kann zum Beispiel ein spezialisierter Recherche-Bot erledigen.
- **Reason:** Je nach Problemstellung erarbeitet das KI-System einen passenden Lösungsansatz. LLMs agieren dabei als Reasoning-Engine.
- **Act:** Anschließend löst das KI-System bestimmte Handlungen aus. Technisch geschieht dies über API-Aufrufe an integrierte bzw. im Web verfügbare Systeme und Services.
- **Learn:** Damit das System kontinuierlich dazulernt, ist häufig eine Lernkomponente integriert – zum Beispiel in Form eines KI-Agenten, der auf die Qualitätssicherung der Ergebnisse spezialisiert ist.

Kerncharakteristiken von Agentic AI im Überblick

- **Autonomie & eigenständige Entscheidungsfindung:** Per Definition setzt Agentic AI auf einen höheren Grad von Autonomie. KI-Bots treffen zu einem gewissen Grad eigene Entscheidungen, ohne die Einbeziehung menschlicher Nutzer.
- **Handlungsfähigkeit durch Zugriff auf externe Systeme:** Agentic AI erweitert den Radius der KI durch die Integration mit weiteren Systemen. Die KI-Bots sind in der Lage, Datenbanken zu durchstöbern und Workflows anzustoßen.
- **Proaktives Agieren:** Die Kombination aus eigener Entscheidungsfindung und dem Initiieren von Handlungen ermöglicht ein proaktives Agieren während des gesamten Lebenszyklus eines Bots.
- **„Meta-KI"/Kaskaden-Prinzip:** Agentic AI lebt von der Kombination mehrerer KI-Modelle und -Fähigkeiten. Typischerweise steuert eine übergeordnete KI mehrere KI-Agenten, die jeweils Spezialaufgaben erfüllen.

Risiken und Herausforderungen beim Einsatz von KI-Agenten

Bereits vor 15 Jahren wurde mit dem Flash Crash von 2010 deutlich, welches Risiko von autonom agierenden KI-Systemen ausgeht. Innerhalb von 36 Minuten verloren führende US-Aktienindizes Billionen US-Dollar an Marktwert. Der Grund dafür waren autonom agierende Algorithmen im Hochfrequenzhandel.

Heute sind die Autonomierisiken von künstlicher Intelligenz noch viel weitreichender. Wenn autonome KI die falschen Entscheidungen trifft, stehen im allerschlimmsten Fall Menschenleben auf dem Spiel – sei es beim autonomen Fahren oder in medizinischen Anwendungen. Im Unternehmenskontext können bereits geringfügige Fehler zu großen materiellen und immateriellen Schäden führen.

Kleine Fehler mit großen Auswirkungen

GenAI-Modelle machen Fehler. Sie halluzinieren, treffen Falschaussagen und verrechnen sich. Auch wenn die Modelle sukzessive besser werden, ist ein kritischer Umgang mit ihrem Output unerlässlich. In Agentic AI-Systemen, die auf diesen Modellen aufbauen, können solche Fehler unentdeckt bleiben und gravierende Konsequenzen nach sich ziehen. Fehlerhafte Zwischenergebnisse führen zu falschen Entscheidungen, die wiederum unangebrachte Aktionen auslösen. Das Problem potenziert sich. Kleine Fehler in frühen Phasen der Verarbeitungskette eines Agentic AI-Systems können fatale Auswirkungen haben.

Agentic AI-Systeme erfordern per se keine Interaktion. Wenn sie nicht mit einer Art Haltbarkeitsdatum versehen werden, können sie ihre Ziele über lange Zeiträume verfolgen. Schwer vorhersehbare Konsequenzen ergeben sich vor allem dann, wenn autonome AI-Bots mit gegensätzlichen Zielen aufeinanderprallen. Der Harvard-Jurist Jonathan Zittrain vergleicht die Situation mit Satelliten, die ins Orbit geschossen werden, in Vergessenheit geraten und irgendwann kollidieren.

Unklarer rechtlicher Rahmen

Agentic AI zeigt deutlich, wie schwierig es für die Gesetzgebung ist, mit den rasanten Entwicklungen im KI-Umfeld Schritt zu halten. Die Begriffe „Agent“ und „agentic“ kommen in aktuellen Rahmenwerken wie dem EU AI Act, ISO 42001 oder dem NIST AI Risk Management Framework nicht vor. Die Definition des AI Act für „KI-Systeme“ trifft jedoch auch für Agentic AI zu. Für sogenannte Hochrisiko-KI-Systeme ist eine menschliche Aufsicht vorgeschrieben. Hier liegt ein großes Konfliktpotenzial, da potenziell nicht aufeinander abgestimmte Ansätze aufeinanderprallen. Wie sieht eine verantwortungsvolle Aufsicht aus, wenn das System autonom – also gerade weitgehend ohne menschliche Eingriffe – arbeiten soll?

Unklar ist auch, unter welchen Umständen ein Agentic AI-System als Hochrisikosystem einzustufen ist. Hier sind u.a. auch Aspekte wie der Aktionsradius und die Zugriffsrechte des Systems zu berücksichtigen. Je nach Einzelfall kann das Risikopotenzial durchaus unterschiedlich eingeschätzt werden.

Es ist davon auszugehen, dass bestehende Regelwerke weiterentwickelt werden, um mehr Klarheit im Umgang mit agentenbasierten KI-Systemen zu schaffen.

Schadensansprüche – wenn der AI-Bot zu viel verspricht …

Fehler autonomer KI-Systeme können zu finanziellen Verlusten führen. Diese Erfahrung musste beispielsweise die Fluggesellschaft Air Canada machen. Das Unternehmen hatte im Kundenservice einen Chatbot eingesetzt, der Zugriff auf die gesamte Website des Unternehmens hatte und eigenständig Kundenanfragen beantworten sollte. Das Problem: Der Bot verstieß gegen die Unternehmensrichtlinien und legte diese großzügiger aus als vorgesehen. Die Fluggesellschaft sah sich schließlich mit Rückerstattungen konfrontiert, die sie nach den Richtlinien ausgeschlossen hatte. Ein Schiedsgericht entschied, dass Air Canada für den Fehler des Bots aufkommen muss.

Technische Risiken

- **Unkontrollierte „Kaskadenfehler“:** Weil Agentic AI aufeinander aufbauende Aktionen ausführt, kann ein kleiner Fehler in einem Teilschritt (z. B. ein Halluzinationsfehler in einem LLM) zu immer größeren Fehleinschätzungen führen. Das Endergebnis kann so stark vom beabsichtigten Use Case abweichen, dass irreversible Schäden entstehen – z. B. falsche Bestellungen oder fehlerhafte Systemkonfigurationen.
- **Falsche Priorisierung bei Konfliktsituationen:** Agentic AI-Systeme sind häufig darauf ausgelegt, eigenständig Entscheidungen zu treffen. Wenn mehrere Ziele und Randbedingungen gleichzeitig gelten, kann die KI Prioritäten falsch interpretieren (z. B. Wirtschaftlichkeit vs. Sicherheit) und dadurch konfliktbeladene Ergebnisse hervorbringen wie bspw. falsch eingesetzte Ressourcen.
- **Mangelnde Robustheit gegen Angriff oder Ausfall:** Agenten, die eigenständige Handlungsbefugnisse besitzen, können durch gezielte Angriffe (z. B. Manipulation von API-Schnittstellen) erhebliche Schäden verursachen oder selbst unbrauchbar werden, wenn nicht entsprechende Sicherheitsmechanismen integriert sind.

Rechtliche Risiken

- **Haftungsfragen bei autonomen Fehlentscheidungen:** Je mehr Autonomie ein Agentic-AI-System besitzt, desto weniger ist klar, wer im Schadensfall tatsächlich haftet. Ist es der Betreiber, der Hersteller oder sogar der Nutzer? Rechtsstreitigkeiten und Imageschäden können die Folge sein.
- **Datenschutzverstöße bei automatisierten Datenabfragen:** Agenten, die eigenständig auf externe Systeme zugreifen, könnten unabsichtlich sensible Daten speichern oder teilen. Ohne eine klare Datenfreigabestrategie und -kontrolle drohen Verstöße gegen Datenschutzregeln wie die DSGVO (bzw. andere Datenschutzvorgaben), die zu hohen Strafzahlungen führen können.
- **Unklare Einordnung unter neue oder künftige Gesetzgebung:** Agentic AI ist in aktuellen Regulierungen wie dem EU AI Act nicht explizit definiert. Unternehmen laufen Gefahr, dass ihre KI-Systeme nachträglich als „Hochrisiko“ eingestuft und strikteren Auflagen unterworfen werden, ohne dass entsprechende Vorkehrungen zuvor getroffen wurden.

Organisatorische Risiken

- **Unklare Entscheidungs- und Verantwortungsstrukturen:** Wenn das Unternehmen nicht klar definiert, wer über die Einführung und Überwachung von Agentic AI entscheidet, entsteht schnell ein „Graubereich“. Dies kann zu Verzögerungen, Fehlentscheidungen und mangelnder Adoption führen, da wesentliche Leitplanken fehlen.
- **Fehlendes Fachwissen bei Mitarbeitenden:** Agentic AI erfordert spezielle Kompetenzen im Umgang mit autonomen Systemen. Fehlen Schulungen zu AI Literacy und zu notwendigen Notfallmaßnahmen, können Mitarbeitende bei unerwartetem Verhalten des Systems nicht angemessen reagieren.
- **Überschreiten der Kompetenzen durch „Selbstorganisation“:** In komplexen Organisationen kann ein KI-Agent – aufgrund großer Handlungsfreiheit – Prozesse verändern oder automatisierte Entscheidungen treffen, die nicht ausreichend mit den betroffenen Fachbereichen abgestimmt sind. Das führt zu ineffizienten Abläufen und Widerständen im Unternehmen.

Ethische Risiken

- **Ungewollte Diskriminierung oder Benachteiligung:** Agenten, die selbstständig Entscheidungen (z. B. Kreditvergabe, Personalvorauswahl) treffen, können vorhandene Vorurteile (Bias) verstärken. Das Resultat sind diskriminierende Ergebnisse, die nicht sofort auffallen, weil kein Mensch direkt eingreift.
- **Verlust der menschlichen Aufsicht:** Wenn Agentic AI-Systeme stark selbstständig agieren, entsteht das Risiko, dass Menschen wichtige Entscheidungsmechanismen verlernen oder gar nicht mehr hinterfragen. So kann mittelfristig wichtiges Prozesswissen verloren gehen.

Ansätze für die Weiterentwicklung der AI Governance

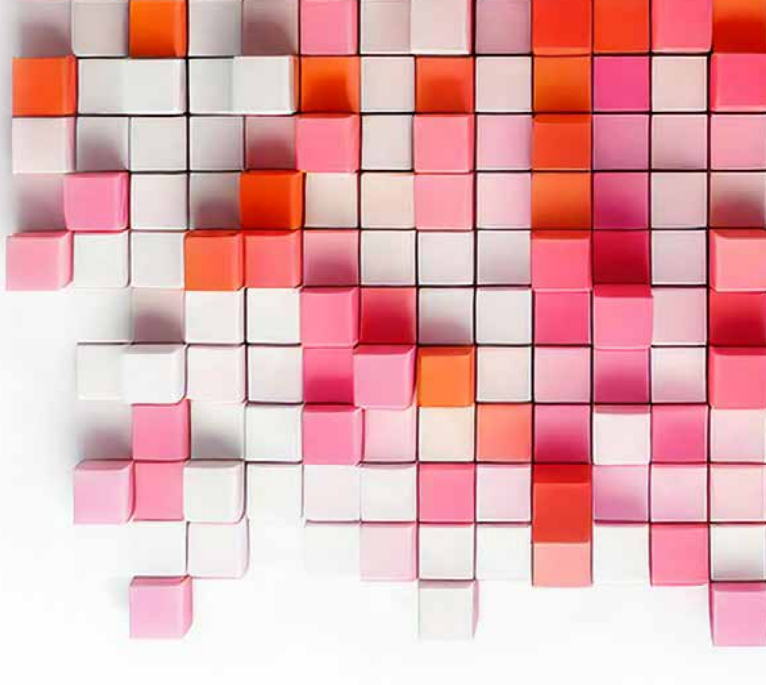
Um die Risiken angemessen zu adressieren und Kontrollverlusten vorzubeugen, muss der Einsatz von Agentic AI-Systemen im Unternehmenskontext mit einer Weiterentwicklung der AI Governance einhergehen. Die gute Nachricht: Es gelten die gleichen Prinzipien wie bei bestehenden Kontrollrahmen für künstliche Intelligenz. Unternehmen können auf gängigen Sicherheits- und Kontrollpraktiken aufbauen, die idealerweise schon implementiert sind.

Darüber hinaus sind neue Kontrollprozesse und Fail-Safes einzuziehen. Eine wesentliche Herausforderung besteht darin, nachvollziehen zu können, wann ein Agentic AI-System Fehler macht. Um ein faires und verantwortungsvolles Handeln der autonomen KI-Bots sicherzustellen, sind zudem ethische Audits sowie Impact-Assessments notwendig. Insgesamt müssen Risiko-Assessments adaptiver, dynamischer und dichter als zuvor gestaltet sein.

Zentrale Kontrollmaßnahmen bei der Einführung von Agentic AI

Die Schwerpunkte einer Erweiterung des Kontrollrahmens sind durch folgende Aspekte bestimmt:

- **Abschaltmechanik bereitstellen und Unterbrechbarkeit von Prozessabläufen ermöglichen:** Auch wenn die Maschine die Arbeit erledigt, muss der Mensch die Kontrolle behalten. Es müssen Mechanismen etabliert werden, um die Agenten zu jeder Zeit stoppen zu können. Ein Forschungsteam von OpenAI schlägt vor, diese Vorgabe bei KI-Agenten immer als primäres Ziel zuverankern: „shut down gracefully when requested by the user“.
- **Aktionsräume der KI klar definieren und einschränken:** KI-Bots sollten nur innerhalb von klar abgesteckten Grenzen operieren. Dafür müssen Freigaben und Zugriffsrechte sorgfältig und so restriktiv wie sinnvoll und möglich konfiguriert sein. Auf welche Daten und Anwendungen haben die Agenten Zugriff? Mit welchen weiteren Systemen sind sie integriert? Außerdem sollten weitere Regeln und Richtlinien wie „Der Agent darf keine Kundendaten mit anderen Kunden teilen“ explizit den Handlungsspielraum des KI-Systems eingrenzen.
- **Menschliche Zustimmung bei kritischen Entscheidungen erforderlich machen:** Was darf das Agentic AI-System entscheiden bzw. tun und was nicht? Es muss klar festgelegt sein, an welchen Punkten der Mensch übernimmt. So könnte zum Beispiel bei einem Buchungssystem der KI-Agent in der Lage sein, alles Nötige vorzubereiten. Er sollte aber – zumindest im ersten Schritt – eher nicht die Berechtigung haben, um Buchungen über einer bestimmten Summe oder Buchungen ohne Stornierungsmöglichkeit durchzuführen.
- **Automatisiertes Monitoring einrichten/erweitern:** Um das Verhalten des Agenten transparent zu machen, ist ein umfangreiches Monitoring unabdingbar. Es hilft dabei, den „Denkprozess“ des Agentic AI-Systems nachvollziehbar zu machen und Fehlerursachen zu identifizieren. Im Rahmen des automatisierten Monitorings kann auch ein parallel laufendes KI-System zum Einsatz kommen.



Pilotprogramm mit Governance-Begleitung – wie die sichere Einführung von Agentic AI gelingt

Die Agentic Governance ist durch eine Reihe von Eigenschaften charakterisiert, die speziell auf die Herausforderung autonomer Abläufe eingeben. Dazu gehören dynamische Entscheidungskorridore ebenso wie Abschaltmechanismen mit Echtzeitalarmen und die Definition von Einstiegspunkten für menschliche Eingriffe.

Um diese wichtigen Grundlagen fall- und praxis-bezogen zu erarbeiten, empfiehlt sich ein Pilotprogramm mit engmaschiger Governance-Begleitung für Agentic AI, das folgende Punkte abdeckt:

Pilotprogramm

Zielsetzung und Scope
Die Aufgaben und der zulässige Aktionsradius sind festzulegen, um das Risiko zu begrenzen. Dies stellt sicher, dass später messbare Ergebnisse vorliegen und das Projekt in einem kontrollierten Rahmen bleibt.

Strukturiertes Testumfeld
Das Pilotprogramm ist in einer kontrollierten Umgebung durchzuführen, z. B. in einer Test- oder Sandboxing-Umgebung. Hier kann das Verhalten des Agenten beobachtet werden, ohne dass er kritische Live-Systeme gefährdet oder Produktivdaten verändert.

Automatisiertes Monitoring und Logging
Ein engmaschiges Monitoring protokolliert alle Entscheidungen und Aktionen des Agenten. Realisiert wird es über ein separates KI-Modul, das ungewöhnliche Verhaltensweisen erkennt und Alarm schlägt. Die gesammelten Daten dienen gleichzeitig als Grundlage für Evaluierungen und Optimierungen.

Menschliche Kontrollpunkte und Abschaltszenarien
Kontrollpunkte sind einzuplanen, an denen menschliche Entscheidungs-träger eingreifen können. Hier gilt es, Abschaltmechanismen festzulegen, sodass der Pilot jederzeit gestoppt werden kann. So lassen sich eventuelle Fehlentwicklungen schnell unterbinden.

Fortlaufendes Learning und Feedback-Schleifen
Nach jeder Testphase erfolgt eine Auswertung, welche Entscheidungen gut und welche weniger gut funktioniert haben. Die Erkenntnisse werden genutzt, um den Agenten kontinuierlich weiterzuentwickeln und die Governance-Richtlinien anzupassen. Auf diese Weise wächst das Pilotprogramm organisch zu einem robusten System heran.

Konkrete Regelungsbereiche für eine Agentic AI Policy

Folgende Aspekte bilden den Rahmen für die Ausarbeitung einer unternehmensspezifischen Agentic AI Policy:

Technische Rahmenbedingungen und Sicherheitsstandards

- **Systemarchitektur und Zugriffsrechte:** Festlegung, welche Schnittstellen (APIs, Datenbanken) Agenten nutzen dürfen und auf welche Daten sie zugreifen können.
- **Fail-Safe-Mechanismen:** Abschalt- und Unterbrechungsmöglichkeiten, um in Notfällen eingreifen zu können.
- **Monitoring und Logging:** Anforderungen an die Protokollierung von Prozessen, wie lange die Daten gespeichert werden und welche Prüfprozesse (z. B. KI-basierte Anomalieerkennung) stattfinden.

Datenmanagement und Datenschutz

- **Datenerhebung und -verarbeitung:** Definition, in welchen Fällen der Agent Daten eigenständig abrufen darf und welche personenbezogenen Daten er verarbeiten darf.
- **Konformität mit Datenschutzvorschriften:** Richtlinien für den Umgang mit sensiblen Daten (z. B. Pseudonymisierung, Verschlüsselung).
- **Transparenz gegenüber Betroffenen:** Beschreibung, wie Nutzer:innen, Kund:innen oder Mitarbeitende informiert werden, wenn ein KI-Agent Entscheidungen trifft, die sie betreffen (Stichwort „Erklärbarkeit“).

Risikomanagement und Qualitätskontrolle

- **Risikobewertung und -klassifizierung:** Kriterien, nach denen Agentic-AI-Anwendungen als „hochrisikoreich“ oder „niedrigrisikoreich“ eingestuft werden.
- **Regelmäßige Audits:** Audit-Prozess (mind. jährlich), bei dem technische, organisatorische und ethische Gesichtspunkte geprüft werden.
- **Krisenmanagement:** Leitlinien für den Umgang mit Fehlfunktionen, Sicherheitsvorfällen oder ethischen Konflikten (z. B. sofortige Unterbrechung des Systems, Meldung an zuständige Behörden).

Ethische und gesellschaftliche Leitplanken

- **Bias und Diskriminierung:** Vorgaben, um systematische Benachteiligungen zu erkennen und zu vermeiden (z. B. regelmäßige Bias-Tests, diverse Trainingsdatensätze).
- **Transparenz und Erklärbarkeit:** Anforderungen an die Nachvollziehbarkeit von Agentic-AI-Systemen, sodass kritische Entscheidungen erklärt werden können.

Schulung und Sensibilisierung

- **Mitarbeitertrainings:** Fortbildungen zu KI-Grundlagen, möglichen Risiken sowie praktischen Eingriffsszenarien (z. B. wie man den Agenten stoppt). Erweiterung der AI Literacy Programme nach dem EU AI Act.
- **Verhaltenskodex:** Code of Conduct für den Umgang mit KI-Agenten, damit Mitarbeitende wissen, was zu tun ist, wenn sie potenzielle Probleme oder Risiken erkennen.

Fazit

Agentic AI eröffnet Unternehmen neue Anwendungsmöglichkeiten für künstliche Intelligenz. Mit Hilfe autonom agierender KI-Bots können sie komplexe, mehrstufige Aufgaben automatisieren und ihre Wettbewerbsfähigkeit steigern. Da das Paradigma technisch auf GenAI-Modellen wie LLMs aufsetzt, ist die Basis in vielen Unternehmen bereits vorhanden.

Der höhere Grad an Autonomie von Agentic AI-Systemen birgt jedoch auch neue Risiken. Fehler der zugrundeliegenden Basismodelle können weitreichende Folgen haben. Die Weiterentwicklung eines robusten AI Governance Frameworks und den damit verbundenen Richtlinien, Prozessen und Kontrollinstrumenten ist unerlässlich.

Um das Potenzial von Agentic AI zu erschließen, empfiehlt es sich, ein internes Agentic AI-Pilotprojekt mit enger Governance-Begleitung zu starten. So lassen sich in einem kontrollierten Rahmen die Fähigkeiten der Agenten ausloten und gleichzeitig eine unternehmensspezifische Agentic AI Policy entwickeln.

Aufgrund der hohen regulatorischen Dynamik muss die AI Governance kontinuierlich weiterentwickelt werden. Der Trend zu einer höheren Autonomie von künstlicher Intelligenz erfordert neue Leitplanken, Standards und Regelwerke. Zudem wächst die Bedeutung von Simulations- und Testwerkzeugen, mit denen das Verhalten von Agentic AI-Systemen auch außerhalb des Produktivbetriebs analysiert werden kann.

Ob Unternehmen Agentic AI gewinnbringend einsetzen können, hängt letztlich eng damit zusammen, inwieweit die Risiken beherrschbar werden. Eine durchdachte Governance und ein stringentes Risikomanagement sind absolut erfolgskritisch.

Ihr Ansprechpartner

Hendrik Reese
Partner und Responsible AI Lead
PwC Deutschland
hendrik.reese@pwc.com

LinkedIn

Über uns

Unsere Mandanten stehen tagtäglich vor vielfältigen Aufgaben, möchten neue Ideen umsetzen und suchen unseren Rat. Sie erwarten, dass wir sie ganzheitlich betreuen und praxisorientierte Lösungen mit größtmöglichem Nutzen entwickeln. Deshalb setzen wir für jeden Mandanten, ob Global Player, Familienunternehmen oder kommunaler Träger, unser gesamtes Potenzial ein: Erfahrung, Branchenkenntnis, Fachwissen, Qualitätsanspruch, Innovationskraft und die Ressourcen unseres Expert:innennetzwerks in 149 Ländern. Besonders wichtig ist uns die vertrauensvolle Zusammenarbeit mit unseren Mandanten, denn je besser wir sie kennen und verstehen, umso gezielter können wir sie unterstützen.

PwC Deutschland. Mehr als 15.000 engagierte Menschen an 20 Standorten. Rund 3,05 Mrd. Euro Gesamtleistung. Führende Wirtschaftsprüfungs- und Beratungsgesellschaft in Deutschland.

www.pwc.de