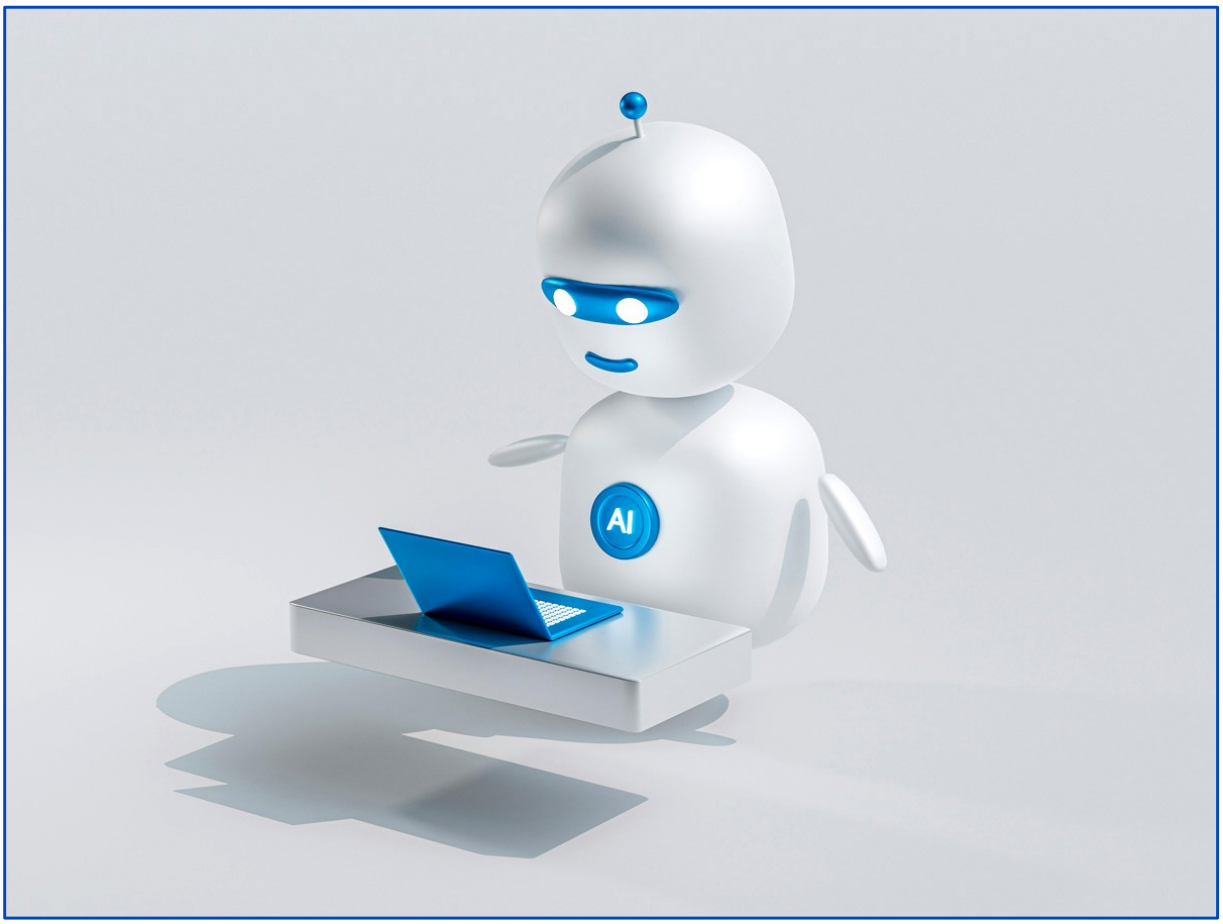
KI-AGENTEN



ZUSAMMENFASSUNG

KI-Agenten bauen auf den Entwicklungen der letzten Jahre im Bereich generativer KI auf, konzentrieren sich jedoch auf die Ausführung von Aktionen in der digitalen oder physischen Welt und nicht auf die Generierung von Inhalten. Durch Kontext-Verständnis, die Nutzung von „Werkzeugen“ und die Fähigkeit zur Planung sind sie in der Lage, komplexe Aufgaben zu lösen und Ziele ohne enge menschliche Aufsicht zu verfolgen. In fortgeschrittenen Einsatzszenarien arbeiten mehrere spezialisierte KI-Agenten zusammen. Von der technologischen Basis der großen Sprachmodelle übernehmen KI-Agenten aber auch deren Nachteile wie Unzuverlässigkeit und mangelnde Fähigkeiten, logisch zu schließen. Bei vielen angepriesenen Szenarien wie automatischer Urlaubsplanung und -buchung stellen sich besonders Fragen nach Haftung und Schutz hochsensibler Daten.

Mit KI-Agenten kommt generative KI vom „Reden ins Tun“

ÜBERBLICK ZUM THEMA

KI-Agenten als jüngste Entwicklung der generativen KI

Im Bereich der Künstlichen Intelligenz (KI) gibt es derzeit unter dem Begriff KI-Agenten neue Entwicklungen, die auf dem Erfolg von generativer KI, die hinter großen Sprachmodellen wie bspw. ChatGPT steht, aufbauen. Während generative KI anfangs primär Inhalte, also Text, Bild, Ton und Video, generierten, erfolgt mit KI-Agenten der nächste Schritt: Agenten sollen gewissermaßen die KI „vom Reden ins Tun" bringen. Während es keine einheitliche Definition von KI-Agenten gibt, so herrscht insofern Einigkeit, dass KI-Agenten – mehr oder weniger eigenständig – Aktionen setzen und Ziele verfolgen können. Dies kann im virtuellen (siehe ***Intelligente soziale Agenten***) oder auch im physischen Raum geschehen (siehe auch ***Cobots*** und ***Humanoide Roboter***). Während Chatbots also erklären können, wie Nutzer:innen Aufgaben erledigen können (diese aber schlussendlich durch Menschen durchgeführt werden müssen), sollen KI-Agenten selber zur Tat schreiten und z. B. Einkäufe tätigen, im Internet nach Informationen suchen oder Termine vereinbaren.

Aktionen setzen statt Inhalte produzieren

Werkzeuge erlauben KI-Agenten Interaktion mit der Umgebung

Ein zentraler Bestandteil von KI-Agenten sind als Werkzeuge bezeichnete Erweiterungen (Kelbert et al., 2024): Während große Sprachmodelle in Reinform nur auf die Information, die beim Training verfügbar war, zugreifen können, machen es Werkzeuge möglich, auch darüber hinaus zu gehen. Eine relativ einfache Form eines Werkzeugs ist der Zugang zur Internetsuche, um etwa auf tagesaktuelle Informationen zugreifen zu können, die beim Training noch nicht berücksichtig werden konnten. Ähnlich sind Werkzeuge, die den Sprachmodellen den Zugriff auf z. B. Unternehmensdaten und -dokumente ermöglichen.

Automatisierte Benutzung von Computern und Webbrowsern

Im letzten Jahr kamen verschiedene KI-Agenten auf den Markt, die einen Computer oder zumindest einen Webbrowser[1] bedienen können. Diese Ansätze ermöglichen es, Arbeitsschritte zu automatisieren, für die früher Programmierkenntnisse und oft auch Zugang zu Programmierschnittstellen notwendig waren. Durch die Fähigkeit, sich sozusagen durch das Internet zu bewegen und auch z. B. Formulare auszufüllen, ist es solchen KI-Agenten möglich, bspw. Hotel- und Flugbuchungen durchzuführen, Produkte zu bestellen und Ähnliches zu tun, was bisher menschlichen Nutzer:innen vorbehalten war.

KI-Agenten zur Lösung komplexer Probleme

Werden mehrere Werkzeuge kombiniert, versprechen KI-Agenten, so komplexe Aufgaben wie die Buchung einer kompletten Urlaubsreise zu automatisieren – zu günstigen Konditionen, in einem terminfreien Zeitraum (durch Kalenderzugriff) und mit einer Unterkunft, die den persönlichen Vorlieben entspricht (Heikkilä 2025). Ein anderer Anwendungsfall sind automatisierte Recherchen im Internet und teilweise in wissenschaftlichen Datenbanken, genannt deep research.[2] Auch in der Softwareentwicklung sollen Agenten zunehmend eigenständig Pro-

[1] Z. B. Operator von OpenAI (*openai.com/index/introducing-operator/*), AI Computer Use von Anthropic (*anthropic.com/news/3-5-models-and-computer-use*) und Googles DeepMinds Mariner (*deepmind.google/technologies/project-mariner/*).

[2] Sowohl OpenAI als auch Google bieten unter diesem Namen Recherche-Assistenten an: *openai.com/index/introducing-deep-research/* und *blog.google/products/gemini/google-gemini-deep-research/*.

grammieraufgaben übernehmen. Gerüchten zufolge plant OpenAI sogar einen KI-Agenten, der umfangreiche Forschungsarbeiten, Datenanalysen und mehr durchführen können soll – und zwar in einer Qualität, die Doktorand:innen ähneln soll (Edwards 2025). Teil der Gerüchte ist aber auch, dass dieser KI-Agent 20.000 US-Dollar pro Monat kosten soll.

Verstehen des Kontexts essentiell

Die erfolgreiche Realisierung von KI-Agenten ist allerdings voraussetzungsreich: Einerseits muss dem KI-Agent beigebracht werden, was seine „Umgebung" ist und wie darauf zugegriffen werden kann. Das heißt, „er" muss „wissen", welche Werkzeuge zur Verfügung stehen und welchen Zweck sie erfüllen – er muss also „verstehen", in welchem Kontext er sich bewegt. Um etwa einen Termin zu organisieren, muss die KI „wissen", ob und wie sie auf den persönlichen Kalender zugreifen kann und wann das sinnvoll ist. Zu diesem Zweck hat etwa Anthropic im November 2024 das Model Context Protocol (MCP) vorgestellt,[3] das den KIs des Unternehmens einen standardisierten Zugang zu verschiedenen Datenquellen (Datenbanken, Dateien oder auch Google Drive) erlaubt.

Vorausschauendes Planen erforderlich, um definiertes Ziel zu erreichen

Um ein Ziel eigenständig verfolgen zu können, müssen KI-Agenten außerdem in der Lage sein, einen Plan auszuarbeiten, umzusetzen und gegebenenfalls anzupassen (Wang et al. 2025). Erst dann ist es möglich, dem KI-Agenten ein relativ vages Ziel vorzugeben, ohne die genaue Ausführung Schritt für Schritt vorzuschreiben. Wird diese Fähigkeit erreicht, versprechen KI-Agenten eine echte Erleichterung für viele Aufgaben zu werden.

Kollaboration mehrerer KI-Agenten zur Erreichung komplexer Ziele

Planende KI-Agenten mit Zugriff auf den Kontext per Werkzeug sind aber nur der erste Schritt der aktuellen Entwicklungen. Der nächste ist, KI-Agenten miteinander kollaborieren zu lassen: So können einzelne KI-Agenten für spezialisierte Aufgaben maßgeschneidert werden, aber auch mit anderen zusammenarbeiten, um komplexere Aufgaben in dynamischen Umgebungen zu erledigen. Ein positiver Nebeneffekt dieser Kollaborationen ist, dass die einzelnen KI-Agenten hochspezialisiert und damit effizienter sein können. Auch hier gibt es aktuell verschiedene Bestrebungen, diese Interaktionen zu standardisieren, etwa durch das von Google im April 2025 vorgestellte A2A-Protokoll.[4] Wichtige Aspekte dieses Protokolls sind u. a. die Koordinierung der Agenten untereinander und der sichere Austausch von Informationen. Das A2A-Protokoll erlaubt die Interaktion mit verschiedenen anderen Technologien (Datenbanken, SAP etc.) und Plattformen (z. B. PayPal) und wird von verschiedenen Dienstleistungsunternehmen (etwa Deloitte, Accenture) unterstützt. Damit soll es auch möglich werden, KI-Agenten über Unternehmensgrenzen zusammenarbeiten zu lassen.

Steuerung von Maschinen und Robotern

Eine andere Stoßrichtung der Entwicklung von KI-Agenten ist die Beeinflussung der physischen Welt durch Roboter. Microsoft hat etwa das KI-Modell Magma entwickelt, das sowohl Computerprogramme als auch Roboterarme steuern kann (Yang et al. 2025).

---

3 *anthropic.com/news/model-context-protocol*.
4 *developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/*.

Generative KI gibt alten KI-Themen neuen Schub

Prinzipiell sind das Planen, die Interaktion mehrerer Agenten (in Form sog. Multi-Agenten-Systeme) und die Steuerung von Maschinen in der KI-Forschung kein neues Thema. Große Sprachmodelle bieten aber eine neue technologische Basis dafür und das aktuell große Interesse gibt diesen Bestrebungen neuen finanziellen Schub

Große Aktionsmodelle zum Lernen von Verhalten statt Text

Eine für KI-Agenten relevante Weiterentwicklung großer Sprachmodelle sind sogenannte große Aktionsmodelle (Large Action Models). Während große Sprachmodelle darauf spezialisiert sind, Texte zu verstehen und zu generieren (indem sie auf der Basis vorangegangener Textteile vorhersagen, welche Worte als nächstes kommen) sind große Aktionsmodelle auf Aktionen trainiert. Das heißt, sie sind gezielt darauf trainiert, Handlungen im digitalen wie im physischen Raum durchzuführen, indem sie zum Beispiel lernen, welche Elemente in einem Programm klickbar sind und welche Schritte zum gewünschten Ergebnis führen – bspw. das Übertragen von Daten in eine Excel-Tabelle (Wang et al. 2025). Ähnlich wie große Sprachmodelle eine große Menge an Textdaten benötigen, brauchen große Aktionsmodelle umfangreiche Datensätze über Nutzer:innen-Intentionen (um Ziele besser zu verstehen), Kontext und gewünschte Aktionen. Oft brauchen sie außerdem eine Form von Gedächtnis und eine Möglichkeit für Nutzer:innen, Feedback zu geben (Wang et al. 2025).

Größere Nützlichkeit wirft neue Fragen auf

Insgesamt versprechen KI-Agenten, generative KI ein gutes Stück nützlicher zu machen, und sie haben das Potential, gerade Routineaufgaben vollständig zu automatisieren. Gleichzeitig wirft der Anspruch, Aktionen möglichst eigenständig auszuführen, auch einige Fragen auf.

Logik ist keine Stärke generativer KI

Mangelnde Verlässlichkeit von generativer KI als Herausforderung

Erstens bauen auch die neuen KI-Agenten auf der Technologie der großen Sprachmodelle auf. Diese haben das ***grundlegende Problem sogenannter „Halluzinationen“***, die zwar im Laufe der letzten Jahre reduziert, aber dennoch nicht grundsätzlich überwunden wurden. Gerade das Planen über mehrere Schritte hinweg verlangt logisches Denken – nicht gerade eine Stärke der generativen KI (Petrov et al. 2025). Auch in letzter Zeit populär gewordene „denkende“ Modelle (mittels sog. Chain-of-Thought-Ansätzen) scheinen so zu „tun-als-ob“. Diese Modelle „denken“ voraus, bevor sie Nutzer:innen eine Antwort liefern – mit dem Ziel, dabei selber Widersprüchlichkeiten zu erkennen und korrigieren zu können. Forscher:innen haben aber gezeigt, dass oft die Denkschritte vor dem Ausgeben einer Antwort nicht konsistent mit der Antwort sind.[5] Bei KI-Agenten, die auch in der virtuellen oder gar physischen Realität Handlungen setzen wollen, kann diese Unzuverlässigkeit unmittelbarer Probleme verursachen als z. B. bei Chatbots. Bedienen KI-Agenten außerdem Roboter, kann mangelhafte Zuverlässigkeit auch zu Sachschäden und Körperverletzungen führen.

[5] *anthropic.com/research/reasoning-models-dont-say-think*; siehe auch: *nytimes.com/2025/05/05/technology/ai-hallucinations-chatgpt-google.html*.

Zugriff auf sensible Daten birgt Risiken

Zweitens wird die Frage der Privatsphäre und Datensicherheit bei Anwendungsfällen wie z. B. automatischen Buchungen, Terminabstimmungen und Zahlungen besonders relevant: Oft ist für eine sinnvolle Nutzung dieser Anwendungen der Zugriff auf hochpersönliche und sensible Daten (Adressbuch, Terminkalender, E-Mails, Zahlungsdaten) notwendig. Sollten diese Daten in falsche Hände geraten, kann dies zu großen Schäden führen. Immerhin gibt es teilweise bereits Ansätze, das Risiko zu minimieren: Durch das Ausführen von KI (und damit der Datenverarbeitung) am persönlichen Gerät der Nutzerin/des Nutzers (Smartphone, Laptop) bzw. der Unternehmensinfrastruktur ist es zunehmend möglich zu gewährleisten, dass Daten das Gerät bzw. Unternehmen nicht verlassen. Diese Entwicklung wird auch als Edge AI oder AI on Device diskutiert und wird durch auf KI spezialisierte Recheneinheiten auf vielen modernen Geräten ermöglicht.

Edge AI und KI on Device als Fortschritt für Datensicherheit

Fragliche Kontrolle über Anreize für die KI

Ein drittes Risiko ist die Steuerung der Anreize von KI-Agenten: Wie können Nutzer:innen wirklich darauf vertrauen, dass ein KI-Agent z. B. wie angewiesen oder versprochen den günstigsten Flug bucht – und nicht etwa über ein Unternehmen, mit dem der KI-Anbieter geschäftliche Verbindungen pflegt?

KI-Agenten werfen Haftungsfragen auf

Letztlich stellt sich schnell auch die Frage der Haftung: Wer kommt für den Schaden auf, wenn z. B. eine Zahlung per KI-Agent getätigt wurde, die von der/dem Nutzer:in nicht intendiert wurde? Welche Verträge können per KI-Agent überhaupt rechtswirksam abgeschlossen werden?

RELEVANZ DES THEMAS FÜR DAS PARLAMENT UND FÜR ÖSTERREICH

Hohe Entwicklungsdynamik mit Relevanz für Österreich als Technologienehmer

KI-Agenten als Weiterentwicklung generativer KI weisen eine hohe Dynamik auf. Gleichzeitig werfen sie einige Fragen besonders in Bezug auf Regulierung auf. So ist nicht ganz klar, welche Rechtsnormen auf KI-Agenten anwendbar sind. Eindeutig sind sie eine Form von Künstlicher Intelligenz i. S. d. AI-Acts der EU. In der Risikoklassifizierung fallen sie aber vermutlich nicht unter Hochrisiko-KI – zumindest nicht generell (Bostoen & Krämer 2024). Je nach Ausgestaltung der KI-Agenten könnten sie aber unter die verbotenen KI-Praktiken fallen, etwa wenn sie die Situation vulnerabler Personen(gruppen) ausnutzen (Art. 5(b) AI-Act).

AI-Act: Vulnerable Personen bedenken

Virtuelle Assistenten durch Digital Markets Act reguliert

Im EU-Gesetz über digitale Märkte (Digital Markets Act, DMA) sind explizit „virtuelle Assistenten“ als zentraler Plattformdienst definiert, und zwar als „eine Software, die Aufträge, Aufgaben oder Fragen verarbeiten kann, auch aufgrund von Eingaben in Ton-, Bild- und Schriftform, Gesten oder Bewegungen, und die auf der Grundlage dieser Aufträge, Aufgaben oder Fragen den Zugang zu anderen Diensten ermöglicht oder angeschlossene physische Geräte steuert“ (Art. 2(12) DMA). Damit fallen KI-Agenten unter eine ähnliche Regulierung wie Suchmaschinen und Online-Marktplätze. Obwohl der DMA nicht in Hinblick auf KI-Agenten von heute und ihren Fähigkeiten entwickelt wurde, auf Sprachassistenten wie Siri oder Alexa mit vergleichsweise begrenzten Fähigkeiten abzielte, ist hier zumindest ein Grundstein gelegt, um Anbieter von KI-Agenten potentiell als Torwächter (Gatekeeper) – bei entsprechender Marktmacht – zu regulieren.

Hohes Schadenspotential für Einzelne und Unternehmen

Eine zentrale Frage bleibt, ob Nutzer:innen unter der aktuellen Rechtslage ausreichend vor Fehlfunktionen von KI-Agenten geschützt sind und ob die Haftungsfrage für KI-Agenten durch neue Rechtsakte präzisiert werden sollte. Bei Fehlfunktion (auch durch Fehlinterpretation der Nutzer:innenwünsche) können, wie bereits oben ausgeführt, potentiell erhebliche Schäden für den Einzelnen entstehen. Aber auch für Unternehmen stellt sich die Frage, ob Verträge, die durch KI-Agenten abgeschlossen werden, rechtswirksam sind. Hier gibt es daher Klärungsbedarf.

Hohe Dynamik wirft Fragen der Regulierung auf

Tätigwerden der KI-Servicestelle angezeigt

Die anhaltend hohe Dynamik im Bereich KI erschwert es teilweise, einen fundierten Überblick über aktuelle Entwicklungen zu behalten. Besonders im Bereich Haftung und Verbraucher:innenschutz gilt es, rasch einen umfassenden Überblick zu schaffen und mögliche Handlungsoptionen zu identifizieren. Sollten keine neuen Rechtsnormen notwendig sein, so sind mindestens klare Empfehlungen und Leitlinien seitens der KI-Servicestelle – für Unternehmen, KI-Anbieter und Verbraucher:innen – anzuraten.

VORSCHLAG WEITERES VORGEHEN

Vor diesem Hintergrund wäre ein kontinuierliches Monitoring der Entwicklungen im Bereich KI-Agenten bzw. allgemein der KI geboten. Ein erster Schritt in diese Richtung wäre eine Kurzstudie, die die in diesem Text aufgeworfenen Entwicklungsoptionen und Fragen vertieft und darüber hinaus einen solchen Monitoringprozess technisch-juristischer Themen definieren könnte.

ZITIERTE LITERATUR

Bostoen, F., & Krämer, J. (2024). *AI Agents and Ecosystems Contestability*. Centre on Regulation in Europe.

Edwards, B. (2025, 7. März). *What does „PhD-level" AI mean? OpenAI's rumored $20,000 agent plan explained*. Ars Technica. ***arstechnica.com/ai/2025/03/what-does-phd-level-ai-mean-openais-rumored-20000-agent-plan-explained/***.

Heikkilä, M. (2025, 12. Februar). *Sam Altman prophezeit: KI-Agenten könnten deinen Alltag erleichtern – aber wie genau?* t3n. ***t3n.de/news/sam-altman-ki-agenten-alltag-erleichtern-aber-wie-1644963/***.

Kelbert, P., Siebert, J., & Jöckel, L. (2024, 19. März). *Large action models (LAMs), tool learning, function calling and Agents*. Blog des Fraunhofer IESE. ***iese.fraunhofer.de/blog/large-action-models-multi-agents/***.

Petrov, I., et al. (2025). *Proof or Bluff? Evaluating LLMs on 2025 USA Math Olympiad* (arXiv:2503.21934). arXiv. ***doi.org/10.48550/arXiv.2503.21934***.

Wang, L., et al. (2025). *Large Action Models: From Inception to Implementation* (arXiv:2412.10047). arXiv. ***doi.org/10.48550/arXiv.2412.10047***.

Yang, J., et al. (2025). *Magma: A Foundation Model for Multimodal AI Agents* (arXiv:2502.13130). arXiv. ***doi.org/10.48550/arXiv.2502.13130***.