Wenn KI handelt: Vertrauen & Verantwortung in Zeiten von Agentic AI

Ethische Verantwortung als Grundlage autonomer Intelligenz

Wenn KI handelt: Vertrauen & Verantwortung in Zeiten von Agentic AI

1. Wenn KI eigenständig handelt: Grundlagen und ethische Dimension der Agentic AI

Autor
Tobias Kellner
Industry Relations Manager Germany, Google

Während klassische KI-Anwendungen, wie Empfehlungssysteme oder Chatbots, bereits fest etabliert sind, beginnt eine neue Ära: die Ära der Agentic AI. Bevor jedoch über dessen Einsatz, dem Potenzial und den Herausforderungen gesprochen werden kann, ist es entscheidend, die Begrifflichkeiten voneinander abzugrenzen.

Agentic AI (dt. agentische KI) bezeichnet eine neue Entwicklungsrichtung und Denkweise der Künstlichen Intelligenz (KI), in der Systeme nicht nur auf Eingaben reagieren, sondern proaktiv Ziele verfolgen, eigenständig handeln und kontinuierlich lernen. Der Begriff umfasst Technologien, Ansätze und Prinzipien zur Gestaltung von KI-Systemen mit agentenhaften Eigenschaften wie Autonomie, Reasoning und adaptivem Verhalten. Es handelt sich deshalb nicht nur um ein einzelnes System, sondern oft um ein Kollektiv von mehreren sich darin befindenden KI-Agenten, die zusammenarbeiten, um komplexe, übergeordnete Ziele zu erreichen.[1]

KI-Agenten stehen dabei für eine neue Generation intelligenter Softwaresysteme, die mithilfe von KI eigenständig Aufgaben ausführen und dabei im Sinne eines Menschen oder Systems handeln. Ein KI-Agent ist ein autonomes digitales System, das KI nutzt, um Aufgaben eigenverantwortlich zu planen, auszuführen, zu bewerten und kontinuierlich anzupassen. Die Fähigkeit zum Reasoning, also zur eigenständigen, logischen Schlussfolgerung, ist dabei elementar. Ein KI-Agent kann somit komplexe Aufgaben analysieren, daraus Handlungsstrategien ableiten, Entscheidungen treffen und seine Aktionen flexibel an neue Informationen anpassen und das mit möglichst geringer menschlicher Unterstützung. [2] Der BVDW hat hierzu bereits im Juni 2025 ein Definitionspapier veröffentlicht.

Wo Agentic AI im Unternehmen stattfindet

Agentic-AI-Systeme werden zunehmend in den unterschiedlichsten Bereichen der Wertschöpfungskette eingesetzt. Sie sind nicht mehr nur in spezialisierten Technologieabteilungen zu finden, sondern werden integraler Bestandteil des täglichen Geschäftsbetriebs.

Einige Beispiele hierfür sind:

- **Marketing und Vertrieb:**
  Autonome Marketing-Agenten, die selbstständig Kampagnen planen, aussteuern und optimieren, um die Kundenzielgruppe bestmöglich zu erreichen.

- **Produktion und Logistik:**
  Intelligente Roboter in Lagerhäusern, die ihre Routen in Echtzeit anpassen, um Engpässe zu umgehen.

- **Kundenservice:**
  KI-Agenten, die nicht nur Anfragen beantworten, sondern proaktiv Probleme identifizieren und Lösungsschritte einleiten.

Die Integration dieser Systeme kann die Effizienz steigern und neue Geschäftsmodelle ermöglichen. Gleichzeitig führt sie jedoch zu tiefgreifenden Veränderungen in den Rollen der Mitarbeiter*innen und den bestehenden Arbeitsprozessen, wodurch neue ethische Fragen aufgeworfen werden.

1 https://www.bvdw.org/wp-content/uploads/2025/06/Definition_KI_Agenten.pdf, S. 2
2 https://www.bvdw.org/wp-content/uploads/2025/06/Definition_KI_Agenten.pdf, S. 2

Wertekonflikte: Eine neue ethische Dimension

Der höhere Grad an Autonomie ist der Kern der disruptiven Kraft von KI-Agenten und der Agentic AI. Dies ermöglicht eine Effizienz und Skalierung, die bisher undenkbar war. Doch mit dieser gesteigerten Autonomie wächst auch die Komplexität der ethischen Fragen, die sich gestellt werden sollten. Die zentrale Hypothese dieses Whitepapers lautet daher:

> **„Je höher der Autonomiegrad einer KI, desto höher die ethischen Anforderungen an ihren Einsatz."**

Die Implementierung von Agentic AI führt unweigerlich zu neuen Wertekonflikten. Während die Technologie verspricht Prozesse zu optimieren, steht dies oft im Spannungsfeld zu grundlegenden ethischen Prinzipien wie Transparenz, Verantwortung und Fairness.

Ein klassisches Beispiel ist der Zielkonflikt zwischen Effizienz und Transparenz: Ein autonomes KI-System mag eine Aufgabe auf dem effizientesten Weg lösen, dieser Weg ist jedoch unter Umständen für den Menschen nicht nachvollziehbar. Wer trägt die Verantwortung, wenn eine autonome Entscheidung zu einem Schaden führt? Wie wird die Fairness bei der Entscheidungsfindung sichergestellt, wenn die Kriterien der KI nicht offengelegt werden können?

Dieses Whitepaper stützt sich auf die sechs ethischen Prinzipien für die Entwicklung und den Einsatz von KI, die bereits im Dezember 2024 vom BVDW erarbeitet und veröffentlicht wurden.[3] Anhand dieser Prinzipien werden wir die zentralen Konfliktfelder systematisch analysieren und praxisnahe Handlungsempfehlungen ableiten. Ziel ist es, Unternehmen dabei zu unterstützen, Agentic AI verantwortungsvoll und nachhaltig zu implementieren und so Vertrauen, Transparenz und Akzeptanz als zentrale Voraussetzungen für eine erfolgreiche Skalierung zu schaffen.

3 Vgl. https://www.bvdw.org/news-und-publikationen/sechs-ethische-prinzipien-fuer-die-entwicklung-und-den-einsatz-von-ki/

2. Messbare Realität in Bevölkerung und Unternehmen

2.1. Gesellschaftliche Akzeptanz und Nutzung von KI-Agenten

Autorin
Katharina Jäger
Head of Innovation & Technology, BVDW

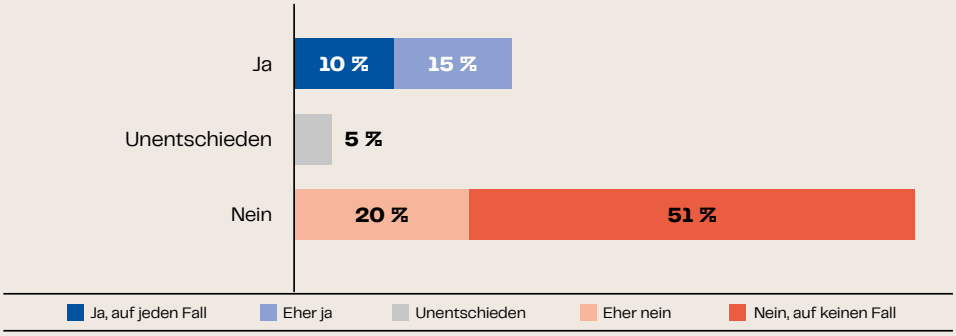
Inmitten schneller technologischer Fortschritte rücken KI-Agenten immer stärker ins Blickfeld, von einer visionären Idee hin zur greifbaren Alltagsanwendung. Aufgaben können erfüllt werden, ohne dass die Menschen hier noch groß eingreifen müssen. Im Kern steht dabei ein Anspruch, der jede digitale Innovation leiten sollte: Technologie soll dem Menschen dienen, nicht umgekehrt. Doch wie bereit ist die Gesellschaft wirklich, solche Entscheidungen aus der eigenen Hand zu geben und einem Algorithmus anzuvertrauen?

Um diese Frage fundiert zu beantworten und den aktuellen Stand der Akzeptanz und Nutzung von KI-Agenten besser einordnen zu können, hat der BVDW über das Meinungsforschungsinstitut Civey eine repräsentative Erhebung durchgeführt. Befragt wurden 2.500 Menschen aus der breiten Bevölkerung zu ihrer grundsätzlichen Offenheit gegenüber KI-Agenten im Alltag. Das zentrale Leitmotiv der Studie: Sind wir bereit, Kontrolle an künstliche Intelligenz zu übergeben und wenn ja, unter welchen Bedingungen?

Die Befragung fragte danach, wie offen die Menschen grundsätzlich dafür sind, Aufgaben vollständig an Künstliche Intelligenz zu übergeben, ohne selbst aktiv eingreifen zu müssen. Dabei ging es um typische Szenarien wie Reisebuchungen oder Produktauswahl.

Die Ergebnisse zeichnen ein klares Bild: Eine klare Mehrheit steht dem Konzept derzeit skeptisch gegenüber. 71 % der Befragten erklärten, dass sie sich nicht vorstellen können, Aufgaben vollständig von einer KI erledigen zu lassen. Innerhalb dieser Gruppe äußerten sich 51 % sogar mit einem eindeutigen „auf keinen Fall“. Nur 25 % zeigen sich grundsätzlich offen und lediglich 10 % befürworten den Einsatz „auf jeden Fall“. Diese Spanne verdeutlicht die Diskrepanz zwischen technologischem Fortschritt und gesellschaftlicher Akzeptanz.

Könnten Sie sich vorstellen, dass Künstliche Intelligenz künftig bestimmte Aufgaben direkt für Sie übernimmt, ohne dass Sie selbst noch aktiv eingreifen müssen (z. B. Reisebuchung, Produktauswahl)?



Stat. Fehler Gesamtergebnis: 3,7% | Stichprobengröße: 2.504 | Befragungszeitraum: 02.07.25 – 03.07.25

Civey

Die Zahlen spiegeln eine deutliche Zurückhaltung wider: Die Bereitschaft, Entscheidungen an KI-Systeme zu delegieren, ist in der Bevölkerung aktuell noch gering. Es geht um zentrale Fragen zu Kontrolle, Vertrauen und digitaler Mündigkeit. Viele Menschen empfinden einen Verlust an Kontrolle, wenn KI-Agenten eigenständig handeln. Gleichzeitig fehlt häufig das Vertrauen in die Transparenz und Zuverlässigkeit der Systeme, ebenso wie in die Akteure, die sie bereitstellen. Auch ein Mangel an digitaler Kompetenz und ein unklarer rechtlicher Rahmen tragen zur Zurückhaltung bei.

Die Studie in der Bevölkerung macht deutlich: Für eine breite gesellschaftliche Akzeptanz von KI-Agenten braucht es weit mehr als technische Perfektion. Essenziell sind umfassende Aufklärungsmaßnahmen, klare ethische Leitlinien und verbindliche rechtliche Rahmenbedingungen. Erst wenn die Menschen verstehen, wie solche Systeme arbeiten, welche Grenzen sie haben und wie sie kontrolliert werden können, lässt sich das Potenzial in vollem Umfang ausschöpfen.

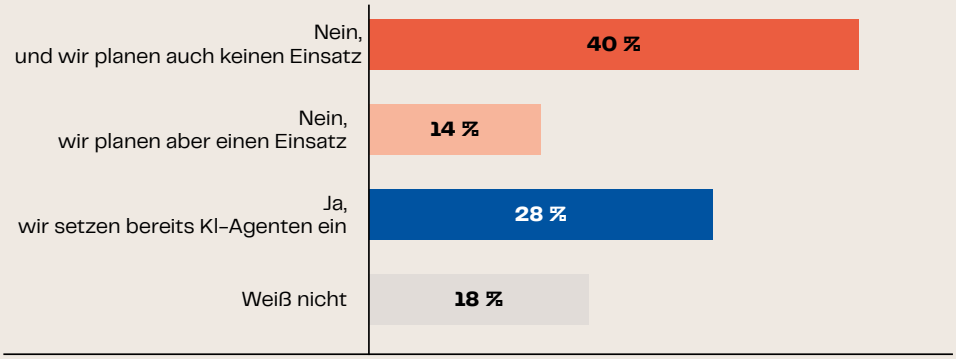
2.2 Einsatz von KI-Agenten in der Privatwirtschaft

Die gesellschaftliche Akzeptanz ist jedoch nur eine Seite der Medaille, ebenso entscheidend ist die Frage, wie offen die Privatwirtschaft für den Einsatz von KI-Agenten ist. Um den aktuellen Status quo zu erfassen, wurden ebenso über das Meinungsforschungsinstitut Civey im Auftrag des BVDW gezielt Entscheidungsträger*innen aus Unternehmen befragt, ob KI-Agenten bereits eingesetzt werden oder ob eine Einführung in naher Zukunft geplant ist.

Unter den knapp 1.000 befragten Unternehmensvertreter*innen ergibt sich ein differenziertes Bild: 28 % gaben an, dass ihr Unternehmen bereits mit KI-Agenten arbeitet. Weitere 14 % planen konkret die Einführung entsprechender Systeme. Zusammengenommen bedeutet dies: Bei mehr als vier von zehn Unternehmen sind KI-Agenten entweder bereits Realität oder stehen unmittelbar vor dem Einsatz.

Gleichzeitig zeigt sich jedoch eine deutliche Zurückhaltung. 40 % der befragten Unternehmen planen derzeit keinen Einsatz solcher Systeme. Weitere 18 % konnten dazu keine klare Aussage machen. Diese Zahlen machen deutlich, dass der Einsatz von KI-Agenten, obwohl technologisch bereits möglich und marktreif, noch weit davon entfernt ist, zum Standard in der Unternehmenspraxis zu werden.

Setzt Ihr Unternehmen bereits KI-Agenten (Selbstlernende Software, die Aufgaben automatisch löst) ein oder planen Sie den Einsatz in naher Zukunft?
Privatwirtschaftliche Entscheidungstragende



Stat. Fehler Gesamtergebnis: 6,8% | Stichprobengröße: 985 | Befragungszeitraum: 02.07.25-10.07.25

BVDW Civey

Während erste Vorreiter also produktiv mit agentenbasierter KI arbeiten, verharrt ein großer Teil des Marktes in der Beobachtungs- oder sogar Ablehnungsphase. Ein zentraler Grund dafür kann die enorme Geschwindigkeit sein, mit der sich KI-Technologien entwickeln. Viele Unternehmen konzentrieren sich aktuell noch auf den Aufbau grundlegender KI-Infrastrukturen, die Integration erster Anwendungen in bestehende Prozesse und die Schaffung organisatorischer Voraussetzungen. KI-Agenten erscheinen in diesem Kontext oft als „zweiter Schritt vor dem ersten", denn ihre produktive Einführung scheitert vielerorts noch an fehlenden technischen, prozessualen und kulturellen Grundlagen.

Hinzu kommt ebenso die ethische Dimension, die für viele Entscheidende von wachsender Bedeutung ist. Fragen zu Datensouveränität, algorithmischer Fairness, Transparenz und der Verantwortung bei automatisierten Entscheidungen beeinflussen die Bereitschaft zur Einführung ebenso stark wie technologische Voraussetzungen. Unternehmen stehen damit nicht nur vor einer technologischen und organisatorischen, sondern auch vor einer wertebasierten Herausforderung.

3. „Je höher der Autonomiegrad einer KI, desto höher die ethischen Anforderungen an ihren Einsatz.“

Autor
Felix von Roesgen
ehem. Consultant Corporate Responsibility, ifok

Mit dem Aufkommen von Agentic AI rücken also neue ethische Fragestellungen in den Vordergrund, allen voran die Frage nach der letztendlichen Verantwortung für die Folgen eigenständiger Entscheidungen. Gleichzeitig bleibt die Notwendigkeit bestehen, diesen Herausforderungen mit einem klaren und robusten Wertegerüst zu begegnen.

Die vom BVDW veröffentlichten sechs Prinzipien für eine ethische KI bilden das Fundament für den verantwortungsvollen Einsatz von KI. Diese sind:

- **Fairness:** KI-Systeme sollen niemanden diskriminieren oder benachteiligen.
- **Transparenz:** Die Funktionsweise von KI-Systemen soll einsehbar sein.
- **Erklärbarkeit:** KI-Entscheidungen sollen nachvollziehbar sein.
- **Datenschutz:** Der Schutz personenbezogener Daten soll gewährleistet sein.
- **Sicherheit:** Fehlfunktionen, Manipulationen und Missbrauch sollen verhindert werden.
- **Robustheit:** KI-Systeme sollen auch unter unsicheren Bedingungen zuverlässig funktionieren.[4]

Keiner dieser Werte verliert im Kontext von Agentic AI an Bedeutung, im Gegenteil: Mit zunehmender Eigenständigkeit steigt das Risiko unerwünschter oder unvorhergesehener Konsequenzen, etwa diskriminierender Muster, intransparenter Entscheidungslogiken oder sicherheitskritischer Fehlhandlungen. In einer repräsentativen Civey-Befragung des BVDW aus dem Dezember 2024, der Teil der Veröffentlichung zu der 6 ethischen Leitlinien war, gaben 54 % der Befragten an, dass sie befürchten, KI-Systeme könnten bestimmte Personengruppen diskriminieren. 73 % würden KI-Produkte meiden, wenn deren Funktionsweise nicht transparent ist und 86 % halten Nachvollziehbarkeit für entscheidend, um Vertrauen in KI-basierte Entscheidungen aufzubauen. Darüber hinaus bewerten rund 90 % den Schutz personenbezogener Daten als wichtig oder sehr wichtig, und 86 % legen besonderen Wert auf die Sicherheit und Zuverlässigkeit solcher Systeme.

Diese Zahlen verdeutlichen: Fairness, Transparenz, Erklärbarkeit, Datenschutz und Sicherheit sind nicht nur ethische Grundpfeiler, sondern unmittelbare Vertrauens- und Wettbewerbsfaktoren. Je höher der Autonomiegrad einer KI, desto schwerer sind Fehlentwicklungen im Nachhinein zu korrigieren und desto größer sind die gesellschaftlichen und ökonomischen Folgen. Verantwortungsvolle KI-Prinzipien werden damit zur zwingenden Voraussetzung für Akzeptanz und wirtschaftlichen Erfolg von Agentic AI. Es lohnt sich daher, diese Prinzipien im Kontext autonomer Systeme besonders sorgfältig zu verankern.

Die bereits vorgestellte zentrale These dieses Paper **„Je höher der Autonomiegrad einer KI, desto höher die ethischen Anforderungen an ihren Einsatz.“** bildet daher den konzeptionellen Rahmen, weil sich mit wachsender Entscheidungsfreiheit technischer Systeme nicht nur ihre Handlungsspielräume, sondern auch ihre potenziellen Wirkungen und Verantwortungsdimensionen erweitern. Eine agentische KI kann nur dann gesellschaftlich akzeptiert und wirtschaftlich erfolgreich sein, wenn sie auf einem belastbaren Fundament ethischer Prinzipien ruht, das Vertrauen schafft und Risiken begrenzt.

4 https://www.bvdw.org/news-und-publikationen/sechs-ethische-prinzipien-fuer-die-entwicklung-und-den-einsatz-von-ki/

4. Evaluierung von Agentic AI entlang der Prinzipien verantwortungsvoller KI

Autor*innen:
Maike Scholz
Group Compliance – Squad Lead Digital Ethics, Deutsche Telekom
Michael Schaarschmidt
Senior Data Scientist, Deutsche Telekom

Agentic AI unterscheidet sich grundlegend von klassischen, reaktiven KI-Anwendungen oder einfachen AI-Assistenten: Sie markiert einen Paradigmenwechsel. Agentic-AI-Systeme agieren proaktiv, adaptiv und oft in komplexen, dynamischen Umgebungen. Und hierbei ist Agentic AI nicht als monolithisches Konzept zu verstehen, sondern als Spektrum.

Mit dieser neuen Autonomie und Multidimensionalität gehen jedoch auch erhebliche Herausforderungen einher. Die Fähigkeit von Agentic AI, eigenständig und in großem Maßstab zu agieren, verschärft bestehende ethische und regulatorische Fragestellungen und verlangt nach neuen Governance- und Kontrollmechanismen.

Im Folgenden werden die zentralen sechs ethischen Prinzipien des BVDW systematisch analysiert, die spezifischen Herausforderungen von Agentic AI herausgearbeitet und praxisnahe Vorgaben für die Unternehmensrealität abgeleitet werden.

4.1 Fairness

Klassische KI-Systeme können bereits diskriminierende Ergebnisse liefern, wenn Trainingsdaten verzerrt sind. Agentic AI verstärkt dieses Risiko, da autonome Systeme eigenständig Entscheidungen treffen und diese in großem Maßstab umsetzen können – oft ohne unmittelbare menschliche Kontrolle, die erhebliche Auswirkungen auf Individuen und Gruppen haben können – etwa bei der Kreditvergabe, im Recruiting oder bei der Ressourcenverteilung. Verzerrte Trainingsdaten oder fehlerhafte Zieldefinitionen können dazu führen, dass bestimmte Personengruppen systematisch benachteiligt werden. Die besondere Relevanz von Agentic AI liegt darin, dass solche Diskriminierungen nicht nur reproduziert, sondern durch eigenständige Zielverfolgung und emergente Strategien sogar verstärkt und skaliert werden können.

Verzerrungen im Datensatz (Bias)

Agentic-AI-Systeme lernen aus großen Datenmengen. Sind diese Trainingsdaten verzerrt – etwa, weil sie gesellschaftliche Vorurteile oder historische Diskriminierung enthalten –, können diese Verzerrungen nicht nur übernommen, sondern durch die autonome Zielverfolgung und Skalierbarkeit der Agenten potenziell verstärkt werden. Einmal implementierte Verzerrungen können sich in vielen automatisierten Entscheidungen manifestieren und sind schwer zu erkennen und zu beheben, da die Systeme eigenständig und ohne ständige menschliche Kontrolle agieren.

Bias-Varianz-Abhängigkeit und Modellagnostische Bias-Reduktion

Die Reduktion von Verzerrungen im Agentic-AI-System kann zu erhöhter Varianz und reduzierter Generalisierung führen. Deshalb ist bei jeder Anpassung und jedem Bias-Mitigations-Ansatz (z. B. durch Outcome-Tests oder Fairness-Optimierung) das sogenannte Bias-Variance-Tradeoff zu beachten. Der Bias-Variance-Tradeoff beschreibt das Problem, dass ein zu einfaches Modell wichtige Muster in den Daten verpasst, während ein zu komplexes Modell zu stark auf einzelne Daten reagiert, sodass man einen Mittelweg finden muss, um die Fehler insgesamt zu minimieren. Modellagnostische Techniken wie Post-Processing mit Wasserstein-basierten Methoden bieten eine effektive Möglichkeit, Bias nachträglich zu mindern, ohne das Modell neu trainieren zu müssen, und halten dabei die Varianz unter Kontrolle. Post-Processing mit Wasserstein-basierten Methoden bedeutet, dass die Ausgaben eines Modells nachträglich so angepasst werden, dass ihre Verteilung einer gewünschten Zielverteilung möglichst ähnlich wird, wobei die Wasserstein-Distanz als Maß für die minimale Verschiebung dient.

Handlungsempfehlungen

Vor dem produktiven Einsatz sollte ein expliziter Bias-Variance-Report erstellt werden, der die Auswirkungen von Mitigationsverfahren auf Robustheit und Generalisierbarkeit untersucht. Bei der Fairness-Optimierung sind Bayesian-Optimierungsverfahren und Multiplicative Logit-Anpassungen sinnvoll, um die Trade-offs transparent darzustellen. Multiplicative Logit-Anpassungen sind ein Post-Processing-Verfahren, bei dem man den Logit einer Wahrscheinlichkeit mit einem Faktor multipliziert, um die Vorhersagewahrscheinlichkeiten zu kalibrieren, abzuschwächen oder zu verstärken, ohne das Modell neu zu trainieren

Autonome Entscheidungsfindung und die Möglichkeit, zu korrigieren

Agentic AI trifft eigenständig Entscheidungen und ergreift Maßnahmen, oft ohne direkte menschliche Kontrolle. Werden dabei diskriminierende oder unfaire Entscheidungen getroffen, können diese sich schnell und großflächig auswirken, etwa durch automatisierte Ablehnungen oder Priorisierungen in der Personalrekrutierung, Kreditvergabe oder im Kundenservice. Die Korrektur solcher systemischen Diskriminierung ist im Nachhinein schwierig, da die Entscheidungswege oft nicht mehr transparent nachvollziehbar sind.

Handlungsempfehlungen zu Fairness/Diskriminierungsvermeidung

- Der Begriff „Fairness“ muss im Vorfeld für das Unternehmen so definiert werden, dass es regelmäßig technisch validierbar ist und ggf. auch extern auditiert werden kann.
- Durchführung eines zwingenden Bias-Assessments und Prüfung der Trainingsdaten auf Repräsentativität.
- Die Belohnungsfunktionen der Agenten sind so zu gestalten, dass faire Entscheidungen explizit incentiviert werden.
- Verantwortlichkeiten für die Überwachung und Korrektur unfairer Entscheidungen müssen klar geregelt sein
- Bei festgestellter Diskriminierung ist der Agent unverzüglich zu stoppen und zu korrigieren.

Als Grundsatzanforderung Fairness bedeutet dies: Unternehmen müssen sicherstellen, dass KI-Agenten niemanden systematisch benachteiligen und bei Verdacht auf Diskriminierung sofort eingreifen. Hierfür sind im Vorfeld entsprechende Verfahren, Meldekanäle und verantwortliche Personen zu definieren.

4.2 Erklärbarkeit/Transparenz: Nachvollziehbarkeit und Verantwortlichkeit

Die Nachvollziehbarkeit von Entscheidungen ist eine Grundvoraussetzung für Vertrauen und Akzeptanz von KI-Systemen. Bei Agentic AI verschärft sich das Problem der Intransparenz, da Entscheidungen oft in verschachtelten, multidimensionalen Agenten-Systemen getroffen werden und die Entscheidungswege für Außenstehende kaum nachvollziehbar sind. Fehlende Erklärbarkeit erschwert zudem die Klärung von Verantwortlichkeiten im Schadensfall.

Intransparenz bei autonomen Entscheidungen

Mit zunehmender Autonomie und Komplexität der Agentic AI wird es schwieriger, die Entscheidungslogik nachzuvollziehen. Besonders problematisch ist dies, wenn Entscheidungen weitreichende Folgen für Betroffene haben, aber die Gründe nicht offengelegt werden können.

Verantwortungsübernahme und strafrechtliche Relevanz

Die Frage, wer für autonome KI-Entscheidungen haftet, ist komplex und wirkt sich im Regelfall auf den Anbietenden des Systems aus. Im Schadensfall (z. B. Diskriminierung, Datenschutzverstoß) muss also zum einen mit Blick auf die Wertschöpfungskette klar geregelt sein, ob Entwickelnde, Betreibende oder Nutzende die Verantwortung tragen und intern klare Governance-Strukturen verankert sein, auch mit Blick auf rechtliche oder gar strafrechtliche Konsequenzen.

Besondere Relevanz im Kontext von Agentic AI

Gerade bei Agentic AI ist Transparenz essenziell, da die Systeme eigenständig und oft in Multi-Agenten-Architekturen agieren. Die Komplexität erschwert es, Bias oder Fehlverhalten zu erkennen und zu belegen. Fairness und Nachvollziehbarkeit müssen daher technisch und organisatorisch abgesichert werden.

Komplexität und Validierbarkeit

Die Validierung von Fairness und die Nachvollziehbarkeit von Entscheidungen sind bei Agentic AI besonders anspruchsvoll. Es braucht Explainability Layer, die alle Entscheidungswege, Datenquellen und Zwischenschritte dokumentieren.

Explainability für Multi-Agenten- und Wissensgraph-Architekturen

Da Agentic-AI-Entscheidungen oft aus verschachtelten Agenten und interaktiven Wissensgraphen hervorgehen, ist eine unternehmensweite Explainability-Schicht erforderlich. Diese muss sämtliche agentengesteuerte Entscheidungswege, genutzte Datenquellen und alle Änderungen im Wissensgraph mitprotokollieren und so speichern, dass ältere Versionen nicht überschrieben, sondern erhalten bleiben. Im Rahmen der „Agent Card", einer standardisierten Beschreibung eines KI-Agenten, sollen für jeden Agenten auch die durchlaufenen Wissensgraph-Knoten und alle Zugriffsebenen dokumentiert werden. Für Multi-Agenten-Systeme empfiehlt sich zusätzlich ein „Graph Audit Log", der sämtliche Transaktionen und Änderungen nachvollziehbar speichert. Ein Graph Audit Log bezeichnet ein Prüf- und Überwachungsprotokoll, das sicherheitsrelevante und administrative Ereignisse dokumentiert. Es zeigt, wer oder was mit wem oder was in Beziehung steht und wie sich das über die Zeit verändert hat.

Handlungsempfehlungen

- Für jeden Agenten ist eine „Agent Card" mit Zweck, Datenquellen, Zugriffsrechten und Verantwortlichen zu führen.
- Ein Explainability-Layer muss alle relevanten Entscheidungsdaten protokollieren.
- Bei kritischen Entscheidungen sind laienverständliche Erklärungen bereitzustellen.
- Verantwortlichkeiten müssen klar zugeordnet und dokumentiert werden.

Als Grundsatzanforderung Erklärbarkeit/Transparenz bedeutet dies: Es muss jederzeit nachvollziehbar sein, warum ein KI-Agent wie entschieden hat und wer dafür verantwortlich ist.

4.3 Datenschutz

Zunächst muss betrachtet werden, dass die Datenschutz-Grundverordnung (DSGVO/GDPR) nur dann Anwendung findet, wenn personenbezogene Daten verarbeitet werden. Arbeiten agentische KI-Systeme ausschließlich mit Daten, die keinen Bezug zu identifizierten oder identifizierbaren Personen haben (z. B. reine Steuerung von Produktionsanlagen oder Netzwerken), unterliegen diese Systeme nicht den Vorgaben der DSGVO/GDPR. Mit Blick auf die Datenvermeidung sollte ein System also von Beginn an so gestaltet werden, dass bestimmte Prozess-Schritte nicht-personenbezogenen Daten arbeiten. Damit entfällt eine datenschutzrechtliche Betrachtung zumindest in diesem jeweiligen Prozess-Schritt. In diesem Fall kann dieses Kapitel übersprungen werden.

Sollten personenbezogene Daten zur Anwendung kommen, also Daten, die einen Personenbezug zulassen, ist der Schutz der Privatsphäre ein zentrales Anliegen. Die Gefahr besteht insbesondere darin, dass autonome Agenten neue, ursprünglich nicht vorgesehene Nutzungen für Daten entdecken („Function Creep“) oder sensible Informationen ungefiltert weitergeben. Vor diesem Hintergrund ist sicherzustellen, dass der ursprüngliche Verarbeitungszweck technisch und organisatorisch durchgesetzt bleibt, das Verhalten des Systems für Verantwortliche nachvollziehbar und erklärbar ist und Lösch- sowie Speicherbegrenzungspflichten über den gesamten Lebenszyklus eingehalten werden.

Unternehmen müssen prüfen, ob sie vor Inbetriebnahme ein Data-Protection-Impact-Assessment (DPIA) durchführen und dieses mit Blick auf den Lebenszyklus und die ggf. geänderte Kritikalität der Anwendung auch bei jeder Modelländerung aktualisieren. Dies erfordert insbesondere eine vorgelagerte Abbildung der Datenflüsse, eine Protokollierung der vom Agenten geplanten und ausgeführten Handlungen sowie klare menschliche Eingriffsmöglichkeiten, wenn der Agent von der vorgesehenen Aufgabe abweicht.

Soweit agentische Systeme Entscheidungen automatisiert vorbereiten oder treffen, die rechtliche oder vergleichbare Auswirkungen für Betroffene entfalten, sind zudem Anfechtungs- und Opt-out-Mechanismen nach Art. 22 DSGVO vorzusehen.

Datenschutz: Schutz personenbezogener Daten

Umgang mit Trainingsdaten und Kommerzialisierung

Agentic AI wird oft mit großen, teils sensiblen Datensätzen trainiert. Die Frage, wie diese Daten erhoben, gespeichert und kommerziell genutzt werden, ist zentral. Es besteht das Risiko, dass personenbezogene Daten ohne ausreichende Rechtsgrundlage verarbeitet oder weitergegeben werden.

Umfangreiche Datenerhebung und -verarbeitung

Agentic AI kann auf eine Vielzahl von Datenquellen zugreifen und diese in großem Umfang verarbeiten, um ihre Ziele zu erreichen. Dies erhöht das Risiko, dass auch unerwartete oder besonders schützenswerte Daten (z.B. Gesundheitsdaten) verarbeitet werden.

Haftungsfragen bei Datenschutzverletzungen

Wenn Agentic AI autonom datenschutzrelevante Entscheidungen trifft (z. B. Weitergabe von Daten ohne Einwilligung), ist die Frage der Haftung komplex. Unabhängig vom Grad der Autonomie des KI-Systems muss geprüft werden, wer als Verantwortlicher für die Rechtmäßigkeit der Verarbeitung und die Einhaltung der datenschutzrechtlichen Pflichten einsteht.

Einschätzung

Datenschutz ist eine zentrale KI-ethische Herausforderung, aber nicht exklusiv für Agentic AI. Die Risiken sind bei allen datengetriebenen KI-Systemen hoch, potentieren sich je nach Autonomiegrad noch.

Handlungsempfehlungen generell:

- Vor Inbetriebnahme ist ein Data Protection Impact Assessment (DPIA) durchzuführen.
- Klare Regeln für Datenminimierung, Zweckbindung und Löschung
- Agenten dürfen nur die für den Zweck erforderlichen Daten erhalten; alle anderen sind zu anonymisieren.
- Systeme sollten Betroffenenrechte (Auskunft, Löschung) automatisiert bedienen können.
- Verantwortlichkeiten sind vertraglich und organisatorisch klar zu regeln.

Mit Blick auf komplexe Multi-User- und Multi-Agenten-Systeme ist es entscheidend, die systematische Trennung von Agentenrechten und Benutzeranmeldeinformationen (User Credentials) technisch sicherzustellen. Das Risiko von unbefugter Rechteausweitung („Privilege Escalation“) steigt signifikant, wenn Agenten personenbezogene User-Credentials für Datenabfragen nutzen, statt über ein dediziertes, fein abgestuftes Rechte- und Identitätsmanagement zu verfügen.

Handlungsempfehlungen Multi-User- und Multi-Agenten-Systeme:

- Jeder Abfragevorgang muss eindeutig zwischen Agentenrecht und Benutzerkontext trennen.
- Wo mehrere Benutzer mit unterschiedlichen Rechten auf das gleiche System zugreifen, darf der Agent niemals seine Rechte missbrauchen oder Userrechte mischen.
- Kontext-sensible RBAC-Systeme (Role-Based Access Control) sind zu verpflichten; die Autorisierung erfolgt immer explizit und nachvollziehbar.
- Jede Änderung im Berechtigungsmanagement wird auditierbar protokolliert („Permission Matrix Logging“).

Hinweis:
Die Integration solcher Mechanismen ist für die Einhaltung der DSGVO essentiell, insbesondere hinsichtlich Artikel 22 der DSGVO (Recht auf nicht ausschließlich automatisierte Entscheidung).

Als Grundsatzanforderung Datenschutz bedeutet dies: Die Vermeidung der Verarbeitung von personenbezogenen Daten sollte das oberste Ziel sein. Sollte dies nicht möglich sein, müssen Fairness und Diskriminierungsvermeidung priorisiert werden.

4.4 Sicherheit

Agentic-AI-Systeme greifen aktiv in Unternehmensprozesse, IT-Infrastrukturen oder sogar physische Systeme ein. Die Gefahr von Manipulation, Fehlfunktionen oder gezielten Angriffen ist daher besonders hoch. Die Autonomie der Agenten vergrößert die Angriffsfläche erheblich, da ein kompromittierter Teil-Agent Befehle an andere weitergeben oder selbst zum Angreifer werden kann.

Späte Entdeckung von Manipulation/Missbrauch

Da Agentic AI ohne ständige menschliche Kontrolle agiert, können Manipulationen, Fehlfunktionen oder Missbrauch (z. B. durch Prompt Injection, feindliche Angriffe) lange unentdeckt bleiben und großen Schaden anrichten.

Besondere Relevanz im Kontext von Agentic AI

Die Autonomie und Vernetzung der Agenten vergrößern die Angriffsfläche erheblich. Ein kompromittierter Agent kann andere beeinflussen oder kritische Prozesse stören. Die Möglichkeit, dass Agenten selbst zum Angreifer werden („Shadow Behaviour"), ist real.

Handlungsempfehlungen

- Aufbau einer Zero-Trust-Architektur: Jeder Agent erhält nur minimal notwendige Rechte.

Die Empfehlung der **Zero-Trust-Architektur** wird konkretisiert:
Jeder Agent muss mit einer eindeutigen Identität („Agent ID") und einem eigenen, kryptographisch abgesicherten Rechteprofil ausgestattet sein. Die Authentifizierung geschieht fortlaufend und nicht nur bei der Initialisierung. Übergaben von User Credentials sind technisch zu unterbinden; stattdessen erteilen User das Ausführungsrecht explizit pro Aufgabe.

- Kryptographische Absicherung aller Agent-Kommunikation.
- Implementierung von kontinuierlichem Monitoring und Anomalie Erkennung. Penetration Tests für Agentenkommunikation und Graphen sollten Standard im Deployment-Prozess sein.
- Notfallmechanismen („Kill Switch") für die sofortige Deaktivierung bei Verdacht auf Missbrauch

Ergänzung von Notfallmechanismen:
Agents, die eine unerlaubte Rechteausweitung oder Privilege Escalation versuchen, werden automatisch isoliert. Die Notfallabschaltung („Kill Switch") ist auf Agenten-, Netzwerk- und Graph-Ebene vorzusehen.

Als Grundsatzanforderung Sicherheit bedeutet dies: KI-Agenten dürfen nur das tun, wofür sie autorisiert sind und sollte immer einen „Not-Aus-Schalter" oder Workaround geben, um sie im Notfall zu stoppen.

4.5 Robustheit

Agentic-AI-Systeme müssen auch unter widrigen Bedingungen zuverlässig funktionieren. Unvorhergesehene Eingaben, manipulative Angriffe oder Ausfälle von Dritt-Tools können zu Fehlverhalten führen, das sich durch die Autonomie der Agenten schnell systemisch ausbreiten kann. Besonders kritisch ist, dass fehlerhafte Informationen durch selbstständige Tool-Nutzung und Memory-Updates dauerhaft ins System diffundieren können.

Destruktive Eigendynamik (z. B. Finanzmärkte)

Agentic AI kann durch autonome, schnelle Entscheidungen in sensiblen Bereichen wie den Finanzmärkten zu unerwarteter Marktvolatilität führen. Gartner[5] warnt, dass Agentic AI in dynamischen Umgebungen (z. B. Finanzmärkte) unvorhersehbare, destruktive Effekte auslösen kann, etwa durch selbstverstärkende Rückkopplungen oder unerwartete Interaktionen. Fehlerhafte Strategien können sich systemisch ausbreiten.

Interaktion mehrerer Agenten

Wenn mehrere Agenten miteinander agieren, können unvorhergesehene Dynamiken entstehen, etwa Deadlocks, Endlosschleifen oder gegenseitige Verstärkung von Fehlern. Die Robustheit des Gesamtsystems ist dadurch gefährdet.

Besondere Relevanz im Kontext von Agentic AI

Die Fähigkeit zur autonomen Interaktion und Selbstanpassung macht Agentic AI besonders anfällig für systemische Fehler und schwer vorhersehbare Wechselwirkungen.

Gleichmäßige Verteilung der Vorteile

Die Vorteile von Agentic AI sind nicht automatisch gleichmäßig verteilt. Es besteht die Gefahr, dass bestimmte Akteure oder Gruppen überproportional profitieren, während andere benachteiligt werden.

Einschätzung

Die Robustheit ist eine zentrale Herausforderung für Agentic AI, da Fehler oder Angriffe sich durch die Autonomie und Vernetzung schnell ausbreiten können.

Handlungsempfehlungen:

- Vor Produktivsetzung: Adversarial Training und Testen in isolierten Sandboxes.
- Graceful Degradation berücksichtigen: Fällt ein Tool aus, muss der Agent auf Alternativen oder menschliche Kontrolle umschalten.
- Dokumentation sicherstellen: Versionierung und Rollback aller Modelle und Daten.
- Während des Lebenszyklus: Systemische Stresstests vor jedem Release.

Als Grundsatzanforderung Robustheit bedeutet dies: Die Systeme werden so konzipiert, dass Fehlfunktionen schnellstmöglich detektiert werden können, sie auch bei Störungen nicht komplett ausfallen und Fehler je nach Kritikalität schnellstmöglich behoben werden können.

5 When to Use or Not to Use AI Agents, Gartner, June 25, 2025 https://www.gartner.com/doc/reprints?id=1-2LJJB-7KU&ct=250729&st=sb&submissionGuid=929383ea-956a-44b0-9ac4-51febc890438

4.6 Multi-Agenten-Robustheit und Trust Boundaries

Die Interaktion mehrerer Agenten in dynamischen Umgebungen birgt systemische Risiken wie Deadlocks, wo zwei oder mehr Agenten aufeinander warten und keiner mehr agieren kann, Endlosschleifen und gegenseitige Verstärkung von Fehlern, insbesondere im Bereich der Trust Boundaries, der Grenze zwischen Bereichen mit unterschiedlichem Vertrauensniveau innerhalb eines KI-Agenten-Systems.

Handlungsempfehlungen:

- Verwendung von verteilten Trust Boundary Enforcement Policies: Jeder Agent darf nur auf authentisierte, kontextbezogene Graph-Knoten und Datenbereiche zugreifen.
- Implementation von „Trust Auditors“: Regelmäßige Analyse von Agenteninteraktionen auf Missbrauchsanzeichen.
- Adversarial Testing für Multi-Agenten: Sandboxing von Kommunikationsströmen und Rollback-Funktion für Graph-Manipulation.

Praxisrelevante Einbettung in die Unternehmensrealität

Die Umsetzung dieser Vorgaben erfordert eine klare Verankerung entlang der gesamten Wertschöpfungskette: Bereits in der Ideation-Phase sollte ein interdisziplinäres „Ethik-Board“ die Mission und das Risikoprofil des geplanten Agentic-AI-Systems prüfen. In der Design- und Build-Phase sollen dann die technischen und organisatorischen Schutzmaßnahmen konzipiert und implementiert werden. Im Deployment sorgen DevOps-Teams für die sichere Infrastruktur und die Integration von Notfallmechanismen. Während des Betriebs überwachen spezialisierte AI-Ops-Teams kontinuierlich die Einhaltung der definierten Vorgaben. Regelmäßige Audits und ggf. auch externe Prüfungen sollen darüber hinaus sicherstellen, dass die Systeme auch langfristig den ethischen und regulatorischen Anforderungen genügen.

Agentic AI eröffnet neue Chancen, stellt Unternehmen aber auch vor erhebliche ethische und technische Herausforderungen. Nur durch die konsequente Umsetzung von Fairness-, Transparenz-, Erklärbarkeit, - Sicherheits-, Datenschutz- und Robustheitsvorgaben kann das Vertrauen in autonome KI-Systeme gestärkt und ihr Potenzial verantwortungsvoll genutzt und, je nach Kritikalität der Anwendung, das Unternehmen auch vor hohen Strafen geschützt werden. Unternehmen müssen Agentic AI als „hochqualifizierte Robo-Mitarbeiter ohne Kinderstube“ begreifen, die klare Regeln, regelmäßige Kontrollen und jederzeit eine verantwortliche Ansprechperson benötigen.

4.7 Wirkungsmechanismen in Multi-Agenten-Systemen

Zur Einordnung der Fehlerverstärkung in Multi-Agenten-Systemen lassen sich zwei grundlegende Wirkmechanismen unterscheiden, deren Auswirkungen im Folgenden anhand eines vereinfachten Rechenbeispiels illustriert werden:

1. **Unabhängige Fehler aufgrund des „Kettenlängen-Effekts“:**
   In Multi-Agenten-Prozessen steigt die Gesamtfehlerwahrscheinlichkeit mit der Anzahl der beteiligten Agenten, da jeder zusätzliche Verarbeitungsschritt eine weitere potenzielle Fehlerquelle darstellt. Die Fehler sind dabei statistisch unabhängig und verstärken sich nicht gegenseitig.

2. **Systemfehler der gegenseitigen Verstärkung, wenn mindestens ein Agent in der Kette einen relevanten Fehler macht, der sich in der Kette multipliziert:**
   Zusätzliche Risiken entstehen, wenn Agenten über Abhängigkeiten oder Trust Boundaries hinweg interagieren und gemeinsam genutzte Informationen verwenden. In diesem Fall können einzelne Fehler systemische Wirkung entfalten und über mehrere Folgeprozesse hinweg weitergegeben werden.

Das folgende Beispiel verdeutlicht anhand eines vereinfachten E-Mail-Szenarios, wie sich diese beiden Effekte – Kettenlänge und Fehlerverstärkung – quantitativ auf die Gesamtrobustheit eines Multi-Agenten-Systems auswirken:

Beispiel: 1000 E-Mails, jeder Agent macht in 5% der Fälle einen Fehler.

Fall 1 (unabhängig):
Durchläuft eine E-Mail mehrere Agenten nacheinander, erhöht sich mit jedem zusätzlichen Agenten die Wahrscheinlichkeit, dass im Gesamtprozess ein Fehler auftritt. Bei 3 Agenten sind dadurch im Mittel rund 143 von 1000** E-Mails betroffen, bei 10 Agenten schon rund 401 von 1000. Dieser Effekt tritt auf, obwohl die individuelle Fehlerquote jedes einzelnen Agenten lediglich 5 % beträgt. Es handelt sich um einen reinen Kettenlängen-Effekt: je mehr Agenten, desto mehr potenzielle Fehlerstellen.

Fall 2 (gegenseitige Verstärkung):
Entscheidend ist in diesem Fall nicht die reine Länge der Verarbeitungskette, sondern die Tatsache, dass Fehler nicht lokal begrenzt bleiben, sondern sich durch Abhängigkeiten zwischen Agenten verstärken können, etwa wenn Ergebnisse anderer Agenten ungeprüft übernommen oder fehlerhafte Informationen über Trust Boundaries hinweg in gemeinsam genutzte Systeme wie Knowledge Graphen oder CRM-Datensätze geschrieben werden: In diesem Szenario stellen die im ersten Fall ermittelten 143 bzw. 401 fehlerhaften E-Mails lediglich die initialen Schäden dar. Teile davon können zusätzliche E-Mails beeinflussen, bis ein Audit/Rollback greift.

Dieses Beispiel verdeutlicht, dass einzelne Fehler unverhältnismäßig große Folgeschäden nach sich ziehen können und daher auch nur geringe Fehlerquoten bei Multi-Agenten-Systeme unbedingt ernst genommen werden müssen.

Erklärung für die Berechnung der Zahlen:
** Die genannten Werte ergeben sich aus der Annahme, dass eine E-Mail nur dann als korrekt gilt, wenn sie alle beteiligten Agenten fehlerfrei durchläuft. Da jeder Agent eine Fehlerwahrscheinlichkeit von 5 % pro Verarbeitungsschritt aufweist, sinkt die Wahrscheinlichkeit eines fehlerfreien Gesamtdurchlaufs mit jedem zusätzlichen Agenten in der Kette. Die angegebenen Zahlen (143 bzw. 401 von 1000 E-Mails) ergeben sich daraus, dass diese verbleibende Erfolgswahrscheinlichkeit auf ein Volumen von 1000 E-Mails angewendet wird und der Rest als fehlerhaft betrachtet wird.

5. Verantwortung institutionalisieren – Vertrauen durch Schaffung von Struktur

Autorin
Katharina Jäger
Head of Innovation & Technology, BVDW

Aus ethischen Prinzipien müssen überprüfbare Strukturen werden, aus Werten messbare Anforderungen. Nur so kann Agentic AI verantwortungsvoll skaliert werden. Die vorangegangenen Kapitel haben gezeigt, dass technologische Leistungsfähigkeit allein nicht genügt, um Vertrauen und Akzeptanz zu sichern. Erst wenn Unternehmen in der Lage sind, Autonomie systematisch zu steuern, entstehen Rahmenbedingungen, die sowohl Innovation ermöglichen als auch ethische und rechtliche Risiken minimieren.

Die Zukunft agentischer Systeme entscheidet sich nicht in den Algorithmen, sondern in der Governance, die sie umgibt. Vertrauen in autonome KI ist kein Zufallsprodukt, sondern das Ergebnis klarer Regeln, kontinuierlicher Kontrolle und einer Kultur der Verantwortung.

5.1 Autonomie Konsortium – Verantwortung operationalisieren, Autonomie sicher steuern

Autorin
Sofia Soto
AI Consultant – Serviceplan AI Labs, Serviceplan Group

Mit dem Aufkommen von Agentic AI verschiebt sich die Verantwortung in Unternehmen spürbar. Entscheidungen, die bislang in menschlicher Hand lagen, werden zunehmend autonom von Systemen getroffen. Um diese neue Form der Autonomie beherrschbar zu machen, braucht es ein verbindliches Rahmenwerk, das Verantwortung mit der technologischen Entwicklung mitwachsen lässt.

Ein solches Rahmenwerk könnte das **„Autonomie-Konsortium“** bilden. Eine praxisorientierte Struktur, die Unternehmen dabei unterstützt, ethische Prinzipien in operative Steuerungsmechanismen zu überführen. Ziel ist es, sicherzustellen, dass jedes KI-System nur so viel Eigenständigkeit erhält, wie es verantwortbar ist und dass Governance und Aufsicht mit der Autonomie mitwachsen

Die fünf Stufen der Autonomie

1	2	3	4	5
Manuell Der Mensch erledigt die Arbeit, die KI liefert Informationen.	**Unterstützt** Die KI macht Vorschläge, der Mensch entscheidet.	**Semi-autonom** Routineaufgaben werden automatisiert, Ausnahmen werden eskaliert.	**Agentic** Die KI plant mehrstufige Aktionen, nutzt Tools und Memory-Systeme.	**Vollautonom** Die KI handelt ohne menschliches Eingreifen vollständig allein.

Je höher der Autonomiegrad, desto strenger müssen die Kontrollen sein. Unterstützende Systeme müssen nachvollziehbar dokumentieren, wie Entscheidungen zustande kommen. Semi-autonome Systeme dürfen nur innerhalb definierter Grenzen operieren und müssen bei Unsicherheit eskalieren. Agentische Systeme erfordern ein engmaschiges Monitoring, geprüfte Notfallmechanismen („Circuit Breaker“) und eine unveränderliche Revisionsspur sämtlicher Entscheidungen.

Vollautonome Systeme bergen das höchste Risiko. Ihr Einsatz sollte nur nach dokumentierter Datenschutz-Folgenabschätzung und gegebenenfalls in Abstimmung mit Aufsichtsbehörden erfolgen.

Menschliche Aufsicht bleibt unverzichtbar

- **Human-in-the-Loop (HITL):** Der Mensch muss kritische Aktionen freigeben.
- **Human-on-the-Loop (HOTL):** Der Mensch überwacht und kann schnell eingreifen.
- **Human-in-Command (HIC):** Der Mensch setzt Ziele und Vorgaben, das System unterstützt die Ausführung.

Zur Aufnahme eines neuen Use Cases hilft folgender Aufnahmeprozess (Autonomie × Risiko-Aufnahmeformular):

a. Autonomiestufe zuweisen.

b. Den Worst-Case-Schaden schätzen (Niedrig / Mittel / Hoch).

c. Die folgende Regel anwenden:

Hoch → HITL.

Mittel + Semi-Autonomie oder hoher → HOTL.

Niedrig + unterstützt → HIC.

So entsteht ein verbindliches Governance-Modell, das Produktteams unmittelbar Orientierung bietet. Ergänzend sollte jedes Unternehmen ein KI-Governance-Board etablieren, das Deployments pausieren, KI-Sicherheitsbeauftragte ernennen und klare Verantwortlichkeiten festlegen kann (Product Owner, Datenschutzbeauftragte, CISO, Ethikbeauftragte).

Nur durch diese Kombination aus Struktur, Aufsicht und Rechenschaftspflicht lässt sich sicherstellen, dass Agentic AI nicht zum Risiko, sondern zum verantwortungsvoll eingesetzten Erfolgsfaktor wird.

5.2 Verantwortung als Leitprinzip der Autonomie

„Je höher der Autonomiegrad einer KI, desto höher die ethischen Anforderungen an ihren Einsatz."

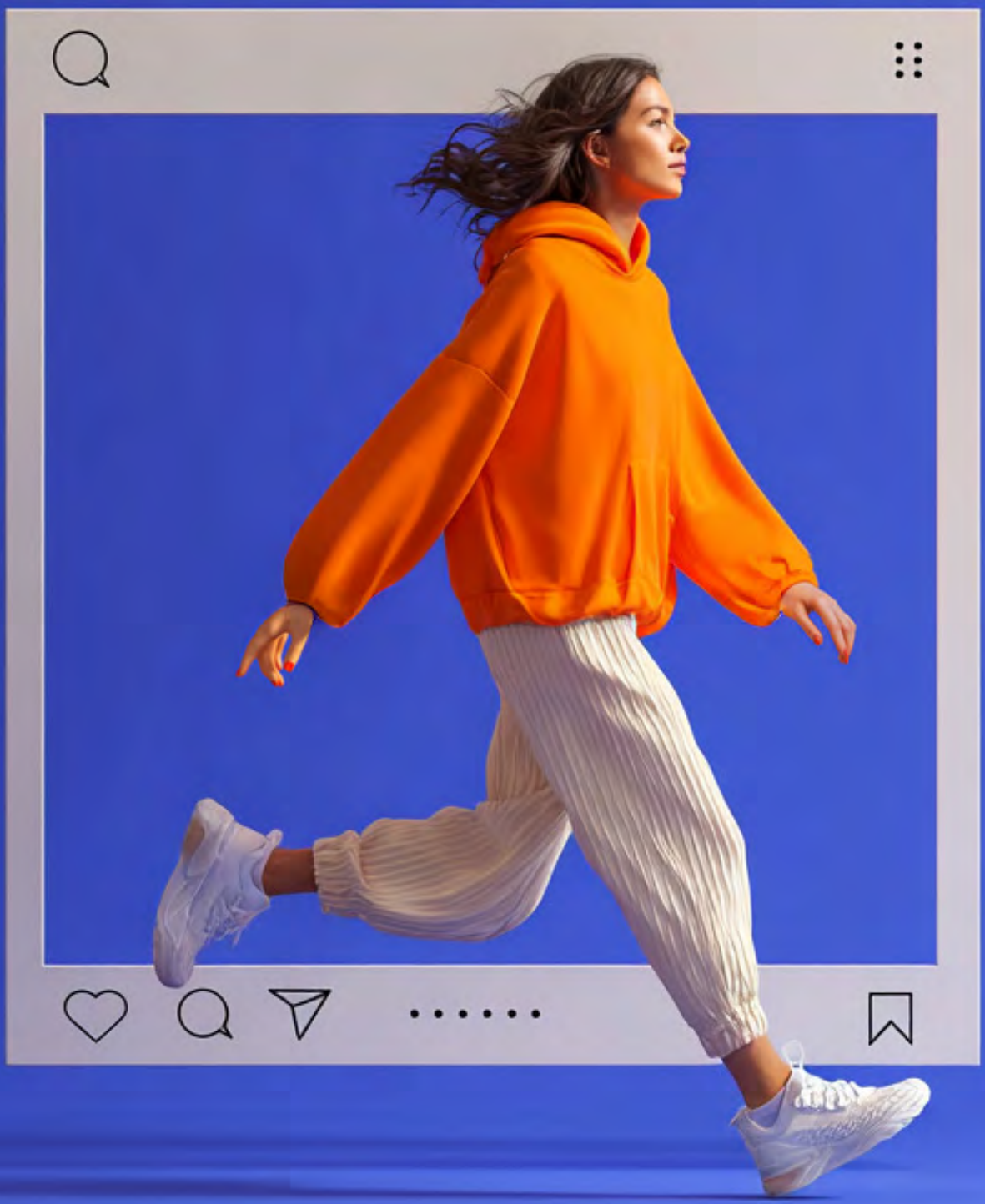
Autorin
Beatriz Bilfinger
Senior Programm Managerin Sustainability & Digital Responsibility, BVDW

Diese Erkenntnis ist nicht nur ein theoretischer Grundsatz, sondern eine Verpflichtung. Mit wachsender Entscheidungsfreiheit technischer Systeme wächst auch ihre Wirkungsmacht und damit die Verantwortung derjenigen, die sie entwickeln, einsetzen und kontrollieren.

Agentic AI kann enorme Potenziale, wie Effizienz, Skalierbarkeit und neue Geschäftsmodelle freisetzen, doch ohne ein tragfähiges ethisches Fundament werden diese Potenziale schnell zu Risiken. Vertrauen, Transparenz und Fairness sind keine Zusatzoptionen, sondern die Voraussetzung für Akzeptanz und wirtschaftlichen Erfolg.

Der BVDW appelliert daher an die Digitale Wirtschaft, gemeinsam Verantwortung zu institutionalisieren. Mit dem Autonomie-Konsortium und weiteren konkreten Handlungsempfehlungen liegen Vorschläge vor, wie sich Prinzipien in überprüfbare Praxis überführen lassen.

Agentic AI braucht Leitplanken, keine Fesseln. Klare Regeln, transparente Prozesse und die feste Überzeugung, dass technologische Stärke erst durch menschliche Verantwortung ihren wahren Wert entfaltet. Es geht hierbei nicht um die Verhinderung technologischer Entwicklung, sondern um ihre Gestaltung, im Einklang mit gesellschaftlichen Werten und unter Wahrung menschlicher Kontrolle. Agentic AI wird die Art, wie Unternehmen arbeiten, grundlegend verändern. Entscheidend ist, ob diese Veränderung von Vertrauen, Verantwortung und ethischer Klarheit begleitet wird. Nur dann kann aus technologischer Autonomie gesellschaftlicher Fortschritt werden.



Autor*innen

Beatriz Bilfinger
Senior Programm Managerin Sustainability & Digital Responsibility,
Bundesverband Digitale Wirtschaft (BVDW) e. V.

Felix von Roesgen
ehem. Consultant Corporate Responsibility, ifok GmbH

Katharina Jäger
Head of Innovation & Technology, Bundesverband Digitale Wirtschaft (BVDW) e. V.

Maike Scholz
Group Compliance – Squad Lead Digital Ethics, Deutsche Telekom AG

Michael Schaarschmidt
Senior Data Scientist, Deutsche Telekom AG

Sofia Soto
AI Consultant – Serviceplan AI Labs, Serviceplan Group

Tobias Kellner
Industry Relations Manager Germany, Google

Bundesverband Digitale Wirtschaft (BVDW) e.V.

Der Bundesverband Digitale Wirtschaft (BVDW) e. V. ist die Interessenvertretung für Unternehmen, die digitale Geschäftsmodelle betreiben oder deren Wertschöpfung auf dem Einsatz digitaler Technologien beruht. Mit seinen Mitgliedern aus der gesamten Digitalen Wirtschaft gestaltet der BVDW bereits heute die Zukunft – durch kreative Lösungen und modernste Technologien. Als Impulsgeber, Wegweiser und Beschleuniger digitaler Geschäftsmodelle setzt der Verband auf faire und klare Regeln und tritt für innovationsfreundliche Rahmenbedingungen ein. Dabei hat der BVDW immer Wirtschaft, Gesellschaft und Umwelt im Blick. Neben der DMEXCO, der führenden Fachmesse für Digitales Marketing und Technologien, und dem Deutschen Digital Award richtet der BVDW auch den CDR-Award, die erste Preisverleihung im DACH-Raum für Digitale Nachhaltigkeit und Verantwortung sowie eine Vielzahl von Fachveranstaltungen aus.
Mehr Informationen finden Sie unter **www.bvdw.org**

Working Group Künstliche Intelligenz

Die gewinnbringende und verantwortungsvolle Nutzung von künstlicher Intelligenz (KI) in der deutschen digitalen Wirtschaft steht im Fokus der Arbeit der Working Group. Ziel ist es, Fragen rund um die Veränderungen der Wertschöpfungskette der digitalen Wirtschaft zu beantworten und Lösungsansätze für die ethischen, sozialen und rechtlichen Herausforderungen durch KI zu bieten, um eine nachhaltige und positive Auswirkung auf die Gesellschaft, Wirtschaft und Umwelt sicherzustellen.

Working Group Digital Responsibility

Wir setzen uns für eine nachhaltige und ethische Nutzung digitaler Lösungen sowie für Rahmenbedingungen für den Einsatz digitaler Technologien ein. Damit die digitale Wirtschaft weiterwachsen kann – und dies in verantwortungsvoller Art und Weise. Dazu hat der BVDW zusammen mit Unternehmen unterschiedlicher Branchen ein europaweites Digital-Responsibility-Framework, die CDR Building Bloxx, erarbeitet.

Impressum

Wenn KI handelt: Vertrauen & Verantwortung in Zeiten von Agentic AI

Erscheinungsort und -datum	Berlin, Januar 2026
Herausgeber	Bundesverband Digitale Wirtschaft (BVDW) e.V. Obentrautstraße 55, 10963 Berlin, +49 30 2062186 - 0, info@bvdw.org, www.bvdw.org
Vorstand gem. § 26 BGB	Carsten Rasner
Präsident	Dirk Freytag
Vizepräsident*innen	Thomas Duhr, Anke Herbener, Corinna Hohenleitner, Dr. Moritz Holzgraefe, Julian Simons, Eva Werle
Kontakt	Katharina Jäger, Leiterin Innovation & Technology, jaeger@bvdw.org Beatriz Bilfinger, Senior Programm Managerin Sustainability & Digital Responsibility, bilfinger@bvdw.org
Vereinsregisternummer	Vereinsregister Düsseldorf VR 8358
Rechtshinweise	Alle in dieser Veröffentlichung enthaltenen Angaben und Informationen wurden vom Bundesverband Digitale Wirtschaft (BVDW) e.V. sorgfältig recherchiert und geprüft. Diese Informationen sind ein Service des Verbandes. Für Richtigkeit, Vollständigkeit und Aktualität können weder der Bundesverband Digitale Wirtschaft (BVDW) e.V. noch die an der Erstellung und Veröffentlichung dieses Werkes beteiligten Unternehmen die Haftung übernehmen. Die Inhalte dieser Veröffentlichung und / oder Verweise auf Inhalte Dritter sind urheberrechtlich geschützt. Jegliche Vervielfältigung von Informationen oder Daten, insbesondere die Verwendung von Texten, Textteilen, Bildmaterial oder sonstigen Inhalten, bedarf der vorherigen Zustimmung durch den Bundesverband Digitale Wirtschaft (BVDW) e.V. bzw. die Rechteinhaber (Dritte).