TRAIBER.NRW
TRANSFORMATION DER AUTOMOTIVE INDUSTRIE IN DER BERGISCHEN REGION

Agentische KI

Frochte, Jörg[1]

Priebe, Sebastian[1]

[1]Hochschule Bochum, Interdisziplinäres Institut für Angewandte KI und Data Science Ruhr (AKIS)

Gefördert durch:

Bundesministerium für Wirtschaft und Energie

aufgrund eines Beschlusses des Deutschen Bundestages

TRAIBER.NRW
TRANSFORMATION DER AUTOMOTIVE INDUSTRIE IN DER BERGISCHEN REGION

Agentische KI

veröffentlicht im Rahmen des Projektes TRAIBER.NRW

PROJEKTINFORMATION

Die fortschreitende Digitalisierung, die Mobilitätswende und der notwendige Wandel zur Klimaneutralität fordern die Automobilzulieferer und die Bergische Region an vielen Stellen. Das Bundesministerium für Wirtschaft und Energie fördert TRAIBER.NRW im Rahmen der Förderbekanntmachung „Transformationsstrategien für Regionen der Fahrzeug- und Zulieferindustrie“ mit 4,1 Mio. EUR bis Ende 2025.

Die Bergische Region umfasst die Städte Remscheid, Solingen, Wuppertal und Düsseldorf sowie den Kreis Mettmann, den Rhein-Kreis-Neuss, den Ennepe-Ruhr-Kreis und den Oberbergischen Kreis. Projektpartner sind automotiveland.nrw e.V., die Bergische Universität Wuppertal, die Gemeinschaftslehrwerkstatt der Industrie von Velbert und Umgebung e.V. (GLW), die Heinrich-Heine-Universität Düsseldorf und die Hochschule Bochum. Das Projekt wurde von den Sozialpartnern der Region initiiert und wird von ihnen maßgeblich unterstützt und begleitet: VBU® Vereinigung Bergischer Unternehmerverbände e.V., Arbeitgeberverband Remscheid und Bergisches Land e.V., IG Metall Velbert, IG Metall Ennepe-Ruhr-Wupper und IG Metall Remscheid-Solingen.

WWW.TRAIBER.NRW

Inhaltsverzeichnis

1 Einleitung 1

1.1 Die Basis agentischer KI - LLMs und generative KI 1

1.2 Definition agentische KI 3

1.3 Wie selbstständig ist ein Agent? Eine Einordnung in Stufen 4

2 Einsatzfelder agentischer KI in Unternehmen 7

2.1 Informations- und Wissensarbeit 8

2.2 Kommunikation und Kundenkontakt 10

2.3 Verwaltung und Koordination 11

2.4 Einkauf und Lieferantenmanagement 12

2.5 IT-Betrieb und Support 13

2.6 Einordnung der Beispiele 14

2.7 Der EU AI Act 16

3 Einführung eines agentischen KI-Systems 16

3.1 Implementation der agentischen KI 18

3.2 Was der Agent wissen muss - die Wissensbasis 18

3.3 KI Werkzeuge 20

3.4 Zu guter Letzt: Prompten 22

4 Sicherheitsaspekte im Umgang mit agentischen Systemen 23

4.1 Handlungsmöglichkeiten und Berechtigungen 24

4.2 Externe Angriffe 24

4.3 Überwachung 25

5 Zusammenfassung 26

1 EINLEITUNG

Die Entwicklung künstlicher Intelligenz wurde seit etwa 2010 deutlich beschleunigt. Maßgeblich dafür waren Fortschritte bei tiefen neuronalen Netzen, leistungsfähigere Grafikprozessoren, große digital verfügbare Datenmengen sowie neue Modellarchitekturen. Während in vielen praktischen Anwendungen zuvor vor allem regelbasierte Systeme, klassische Verfahren des maschinellen Lernens und spezialisierte Modelle eingesetzt wurden, rückten seitdem zunehmend datengetriebene Verfahren des Deep Learning in den Mittelpunkt. Ein sichtbarer Wendepunkt war der Erfolg tiefer neuronaler Netze in der Bilderkennung, insbesondere durch den Einsatz großer Trainingsdaten und GPU-basierter Berechnung [1], [2]. In den folgenden Jahren entstanden neue Modellfamilien für Bild-, Sprach- und Textverarbeitung. Besonders folgenreich war die Transformer-Architektur, die seit 2017 eine zentrale Grundlage vieler moderner Sprachmodelle bildet [3]. Auf dieser Basis entwickelten sich große Sprachmodelle, generative KI-Systeme und zuletzt agentische KI-Systeme, die nicht nur Inhalte erzeugen, sondern Aufgaben über mehrere Schritte hinweg bearbeiten können [4], [5].
Dieses Themenpapier befasst sich mit agentischer KI und ihrer Bedeutung für Unternehmen. Zunächst werden die zugrunde liegenden Technologien eingeordnet und von angrenzenden Begriffen abgegrenzt. Anschließend wird erläutert, wie agentische KI grundsätzlich arbeitet, welche Rolle große Sprachmodelle, Werkzeuge, Datenquellen und menschliche Kontrolle dabei spielen und wo die Grenzen aktueller Systeme liegen. Darauf aufbauend werden mögliche Anwendungsfelder in Unternehmensprozessen beschrieben. Risiken werden nicht nur gesondert betrachtet, sondern an den jeweiligen Einsatzstellen eingeordnet, da sie stark vom konkreten Prozess, vom Autonomiegrad und von den verwendeten Daten abhängen. Betrachtet werden hier KI-Systeme, deren zentrale Steuereinheit ein großes Sprachmodell ist. Daneben existiert ein breites, deutlich älteres Feld spezialisierter KI- und Optimierungsverfahren, die gezielt für das jeweilige Verfahren entwickelt werden.

1.1 DIE BASIS AGENTISCHER KI - LLMS UND GENERATIVE KI

Agentische KI baut auf mehreren Entwicklungen auf, um etwas Neues zu schaffen. Die meisten sind inzwischen mit Chatbots vertraut, deren Kern ein großes Sprachmodell (Large Language Model, LLM) ist. Eine Nachricht (Prompt) wird an das LLM gesendet und dieses bestimmt daraus eine *scheinbar* passende Antwort. Technisch ist ein LLM ein sehr großes neuronales Netz, dessen Milliarden bis hin zu Billionen Parametern beim Training anhand riesiger Textmengen, überwiegend aus dem öffentlichen Internet, eingestellt wurden [4]. Die Trainingsaufgabe ist einfach: das jeweils nächste Wort eines Textes vorherzusagen (genau genommen Wortbestandteile, sogenannte Token). Hierfür muss das Modell jedoch implizit erfassen, welche **Bedeutung** Wörter je nach **Kontext** haben und welche inhaltlichen Zusammenhänge in den Texten stecken, denn welche Fortsetzung wahrscheinlich ist, hängt vom gesamten bisherigen Text ab. Bei der Nutzung geschieht dann genau das: Das Modell berechnet aus dem Prompt für jedes mögliche nächste Wort eine Wahrscheinlichkeit, wählt eines aus,

hängt es an und wiederholt den Vorgang mit dem so verlängerten Text, bis eine vollständige Antwort entstanden ist. Interessanterweise wird dabei bewusst nicht immer das wahrscheinlichste Wort gewählt. Ein Zufallsanteil lässt die Sprache natürlicher wirken und ist ein Grund, warum dieselbe Frage bei mehrfacher Eingabe unterschiedliche Antworten liefern kann. Mit Abschluss seines Trainings gewinnt ein Modell für sich genommen keine neuen Informationen hinzu. Es kann ab diesem Zeitpunkt nur mit dem Trainingsdaten-Stand und was im aktuellen Prompt steht arbeiten, weitere Informationen kann es von sich aus nicht hinzuziehen. Außerdem erzeugt es Antworten, die wahrscheinlich klingen, nicht Antworten, die geprüft wurden. Fehlt Wissen, entstehen mitunter flüssig formulierte, aber inhaltlich falsche Aussagen, sogenannte Halluzinationen.

Im nächsten Schritt wurden LLMs mit einem eigens festgelegten AI-Workflow verknüpft, um es der KI zu ermöglichen, weitere Informationen hinzuzuziehen. So kann zum Beispiel eine vordefinierte Schnittstelle mit einem Kalender dabei helfen, einen Chatbot zu erschaffen, der einem Informationen zu den eigenen Terminen zurückspielen kann. Der Vorteil von individuellen Informationen, die ein LLM nie gelernt haben kann, geht jedoch mit einem gewissen Verlust von Flexibilität einher. Mit einer fest etablierten Schnittstelle müssen nun die Prompts des Nutzers auch thematisch damit verknüpft sein. Es ist natürlich möglich weitere Informationsquellen in diesen AI-Workflow einzubinden und so wieder Flexibilität herzustellen. Dies muss aber immer im Vorhinein von Menschen definiert worden sein [6], [7].

Einen weiteren Schritt in Richtung Autonomie stellen KI-Agenten dar. Hier wird ein LLM nicht nur als Antwortgenerator verwendet, sondern als Teil eines Systems, das Ziele interpretieren, Zwischenschritte planen, Werkzeuge auswählen, Informationen beschaffen und Ergebnisse überprüfen kann. Agentische Systeme verbinden damit sprachliche Fähigkeiten mit Handlungsfähigkeit. In der Forschung wird diese Verbindung häufig als Zusammenspiel von Schlussfolgern und Handeln beschrieben: Das Modell erzeugt nicht nur Text, sondern nutzt Aktionen, um zusätzliche Informationen zu erhalten oder einen Zustand in einer Umgebung zu verändern [8].

Die drei genannten Stufen (Chatbot, AI-Workflow und KI-Agent) sowie eine vierte, der Hintergrund-Agent, unterscheiden sich damit vor allem in einer Frage: Wie viel wird vom Menschen im Voraus festgelegt und wie viel entscheidet das System während der Bearbeitung selbst? Diese Eigenschaft wird im Folgenden als Autonomie bezeichnet. Abb. 1 ordnet die Stufen entlang dieser Eigenschaft und ihrer ungefähren zeitlichen Entwicklung ein. Als vierte Stufe ist dort der Fall ergänzt, dass ein Agent nicht mehr von einem Menschen gestartet wird, sondern durch ein Ereignis, wie einen neu eingehenden Vorgang bei einem System – diese Betriebsart spielt in Kap. 3 eine zentrale Rolle.

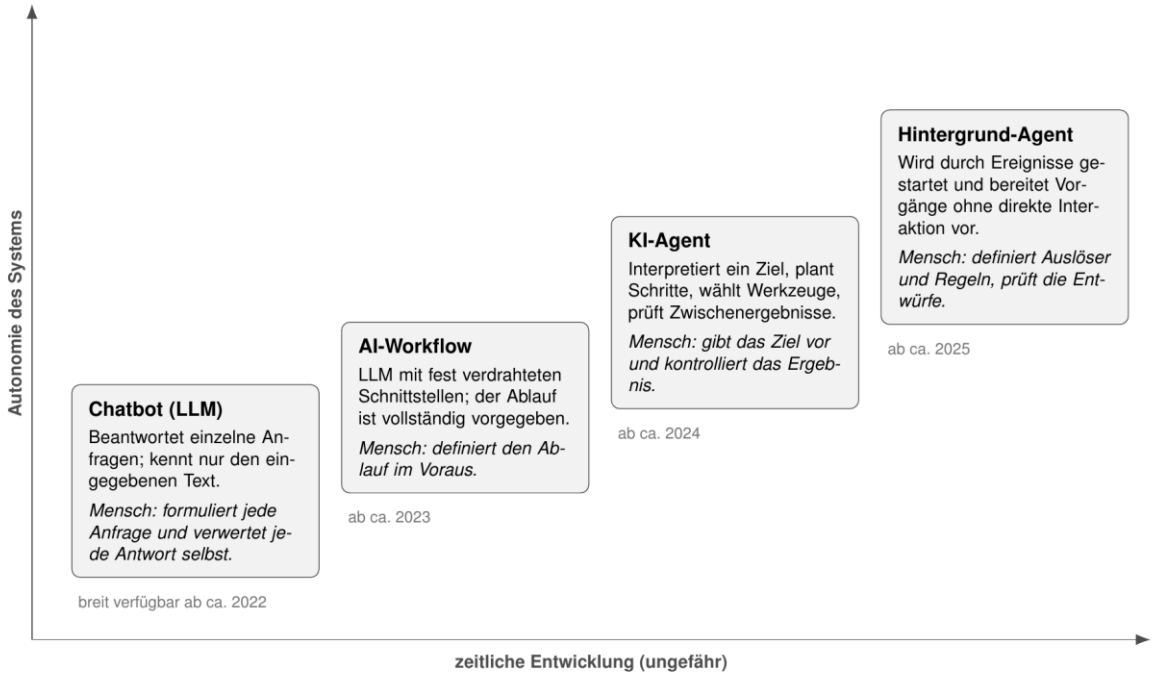

Abbildung 1: Stufen sprachmodellbasierter KI-Systeme geordnet nach ihrer Autonomie und zeitlicher Entstehung. Mit jeder Stufe sinkt die menschliche Beteiligung.

1.2 DEFINITION AGENTISCHE KI

Der klassische Begriff des Agenten bezeichnet ein System, das seine Umgebung über Sensoren wahrnimmt, über Aktoren auf sie einwirkt und dabei versucht, ein vorgegebenes Ziel möglichst gut zu erreichen [9]. Schon ein einfacher Staubsaugroboter ist in diesem Sinn ein klassischer Agent: Er nimmt Hindernisse wahr, entscheidet über seine Route und verändert die Umgebung, indem er sie reinigt. Die Entscheidungslogik kann dabei als fest programmierter Algorithmus vorliegen oder maschinelles Lernen einsetzen; zwischen einfachen und lernenden Agenten verläuft der Übergang fließend. Wesentlich ist der geschlossene Kreislauf aus Wahrnehmen, Entscheiden und Handeln.

Ein Sprachmodell in einem Chatfenster ist, anhand des klassischen Agenten gemessen, ein Grenzfall. Seine gesamte Wahrnehmung besteht aus dem Text, den der Benutzer eingibt, seine einzige Handlung ist die Textantwort. Es kann weder zusätzliche Informationen beschaffen noch etwas in seiner Umgebung verändern, und ohne neue Eingabe geschieht nichts. Agentische KI entsteht, wenn dieses Fenster geöffnet wird: Das Sprachmodell wird zur Entscheidungseinheit eines Agenten, während Werkzeuge die Rolle der Sensorik und Aktorik übernehmen. Ein Datei- oder ein Datenbankzugriff sind im übertragenen Sinn Sensoren, das Ausführen eines Skripts oder das Ablegen eines Dokuments sind im übertragenen Sinne Aktoren. Die Software, die das Sprachmodell auf diese Weise mit Werkzeugen verbindet und die Arbeitsschleife steuert, wird als Harness bezeichnet. Abb. 2 stellt beide Sichtweisen gegenüber.

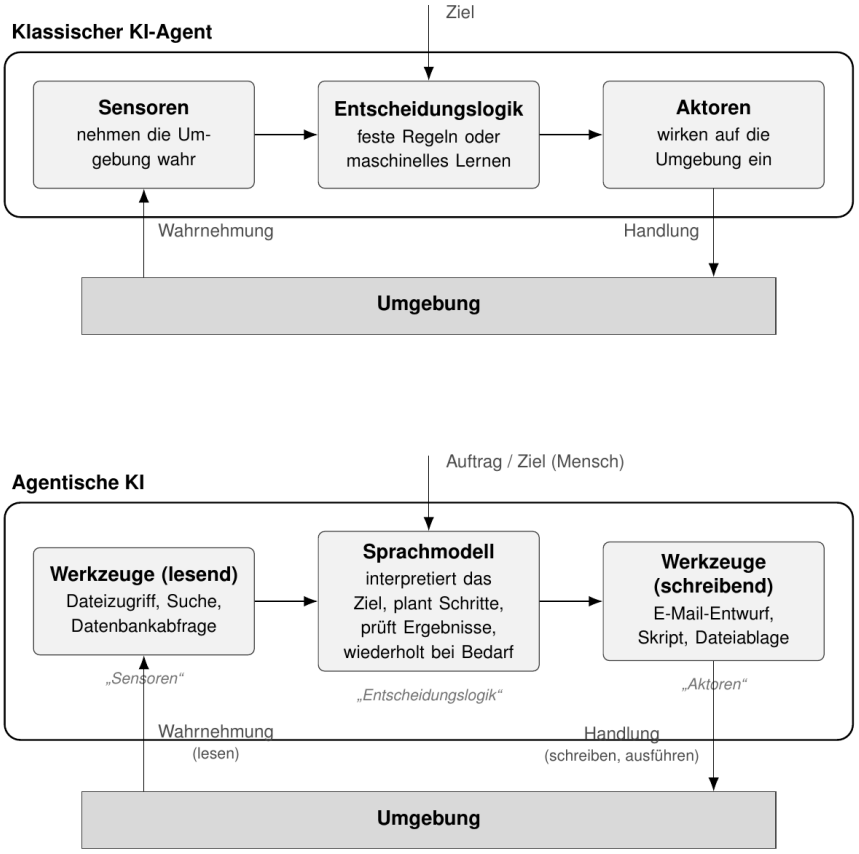

Abbildung 2: Klassischer KI-Agent (oben) und agentische KI (unten) im Vergleich. Die Struktur ist dieselbe: Ein System nimmt seine Umgebung wahr, entscheidet und handelt, um ein Ziel zu erreichen. Bei agentischer KI übernimmt das Sprachmodell die Rolle der Entscheidungslogik, Werkzeuge übernehmen die Rolle von Sensoren und Aktoren. Die Umgebung kann in beiden Fällen nahezu alles sein – von der physischen Welt eines Staubsaugroboters bis zu Dateiablage, Fachanwendungen und Internet; sie unterscheidet die beiden Ansätze nicht grundsätzlich. Auch die Grenze zwischen den Ansätzen ist fließend, da die Entscheidungslogik eines klassischen Agenten ebenfalls maschinelles Lernen nutzen kann.

Aus diesem Blickwinkel lässt sich agentische KI wie folgt charakterisieren: Ein agentisches KI-System verfolgt ein vorgegebenes Ziel weitgehend selbstständig, indem es die Aufgabe in Teilschritte zerlegt, geeignete Werkzeuge auswählt und aufruft, Zwischenergebnisse prüft und bei Bedarf frühere Schritte wiederholt, bis das Ergebnis den Anforderungen entspricht [6], [7], [8]. Generative KI erzeugt aus einem Prompt eigenständig Text, Bilder o. Ä., kann das Ergebnis aber nicht eigenständig weiterverwenden. Es benötigt dafür mindestens einen nichtagentischen Workflow, ein agentisches System kann selbst die nächsten Schritte wie ablegen, versenden oder als Grundlage des nächsten Arbeitsschritts nutzen, planen. Es reagiert damit nicht nur auf einzelne Eingaben, sondern arbeitet einen Auftrag über mehrere Schritte hinweg ab.

1.3 WIE SELBSTSTÄNDIG IST EIN AGENT? EINE EINORDNUNG IN STUFEN

Die Entwicklung in Abb. 1 zeigt, dass KI-Systeme mit jeder Stufe selbstständiger geworden sind. Für die Frage, ob und wie ein Agent im eigenen Unternehmen eingesetzt werden kann, reicht diese Darstellung aber nicht aus. Ein Agent kann sehr selbstständig darüber entscheiden, wann er arbeitet, und trotzdem nur Entwürfe erstellen. Ein anderer wird nur auf Anfrage tätig, darf dann aber Daten in produktiven Systemen ändern. Beide gelten als autonom, die Risiken sind jedoch völlig unterschiedlich. In der Forschung wird Autonomie deshalb schon seit Langem nicht als ein einzelner Wert betrachtet, sondern getrennt nach den Aufgaben, die ein System übernimmt [10], [11], [12]. Für den praktischen Einsatz genügen zwei Fragen: Wer stößt die Arbeit an, und was darf das System ohne Rückfrage bewirken?

Für die erste Frage bietet sich eine Anleihe beim autonomen Fahren an. Dort beschreibt die Norm SAE J3016 Stufen, an denen sich ablesen lässt, wie viel ein Fahrzeug selbst übernimmt und wann der Mensch eingreifen muss [13]. Ähnlich lassen sich agentische Systeme danach einordnen, wie viel Initiative sie haben. Wichtig ist dabei weniger der technische Auslöser als der Auftrag, den das System beim Start erhält. Technisch wird fast jeder Agent entweder durch ein Zeitereignis gestartet, etwa einen Zeitplan, oder durch ein Zustandsereignis, etwa eine neu eingegangene Mail oder ein geändertes Ticket. Auch Systeme, die scheinbar dauerhaft laufen, werden in Wirklichkeit in regelmäßigen Abständen geweckt. Entscheidend ist, ob das System dann eine fest vorgegebene Aufgabe abarbeitet oder selbst entscheidet, ob und was zu tun ist. Tabelle 1 fasst die Stufen zusammen.

Tabelle 1: Stufen der Initiative agentischer KI-Systeme mit Beispielen (Stand Oktober 2026).

Stufe	Beschreibung	Beispiele
0 – auf Zuruf	Ein Mensch startet jede Arbeit und formuliert den Auftrag.	Chat-Assistenten im Browser; Coding-Agenten im Dialog mit dem Entwickler
1 – ausgelöst, fester Auftrag	Ein Zeit- oder Zustandsereignis startet einen vorab festgelegten Auftrag. Das Ereignis liefert nur den konkreten Fall.	Dienstreise-Agent aus Kap. 3; Einkaufsagent, der Mails von Lieferanten auswertet [14]; regelmäßig erstellte Berichte
2 – ausgelöst, offener Prüfauftrag	Das System wird regelmäßig geweckt und entscheidet selbst, ob etwas zu tun ist und was.	persönlicher Assistent OpenClaw mit regelmäßigem “Heartbeat” [15]; Hintergrundagenten, die selbstständig das Gedächtnis eines Agenten pflegen [16]; das inzwischen eingestellte ChatGPT Pulse [17]
3 – erweitert sich selbst	Das System kann im freigegebenen Rahmen wiederverwendbare Skills oder künftige Auslöser anlegen beziehungsweise ändern.	Agenten, die eigene Skills schreiben, teils mit menschlicher Freigabe [18], [19]; frühe Versuche wie BabyAGI, das sich aus einem Oberziel selbst Aufgaben erzeugt [20]

Viele Systeme, die heute in Unternehmen eingesetzt werden, bewegen sich auf Stufe 0 oder Stufe 1. Auch Agenten, die rund um die Uhr auf eingehende Vorgänge reagieren, erhalten in der Regel einen

festen Auftrag. Stufe 2 und Stufe 3 finden sich bisher vor allem bei persönlichen Assistenten und experimentellen Systemen. Dass diese Stufen nicht nur theoretisch riskanter sind, zeigt das Beispiel OpenClaw. Anfang 2026 wurden auf der zugehörigen Plattform für Erweiterungen mehrere Hundert manipulierte Skills gefunden, die Daten der Nutzer abgegriffen haben [21]. Ein offener Auftrag, weitreichende Rechte und die Möglichkeit, sich selbst zu erweitern, verstärken sich hier gegenseitig.
Ein nur lesendes System verändert keine Quelldaten, kann aber vertrauliche Informationen offenlegen oder durch falsche Ausgaben Schaden verursachen. Schreibende Aktionen sind nur dann als reversibel einzustufen, wenn eine geprüfte Wiederherstellung, etwa durch Versionierung, Transaktionen oder Sicherungen, vorgesehen ist. Kombiniert man beide Fragen, ergibt sich die Einordnung in Abb. 3. Je weiter oben und je weiter rechts ein Vorhaben liegt, desto mehr muss vor dem Einsatz festgelegt werden, welche Rechte das System erhält und an welchen Stellen ein Mensch eingebunden wird. Für einen ersten Einsatz eignet sich besonders Stufe 1 in Verbindung mit Entwürfen. Der Agent arbeitet dann selbstständig im Hintergrund, die Wirkung nach außen bleibt aber beim Menschen. Genau so ist das Beispiel der Dienstreise in Kap. 3 aufgebaut.

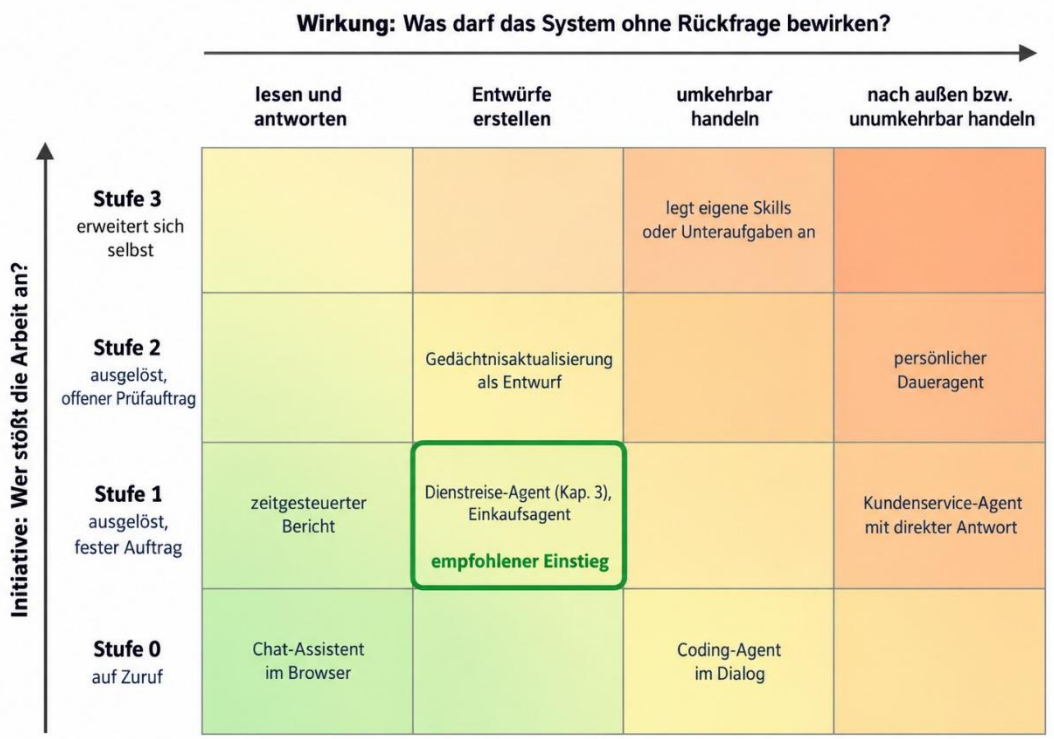

Abbildung 3: Einordnung agentischer Systeme nach Initiative und Wirkung. Die Färbung deutet an, wie stark der Bedarf an vorab festgelegten Grenzen und menschlicher Kontrolle steigt. Eingeordnet wird immer der konkrete Einsatz: Dasselbe Produkt kann je nach Konfiguration in verschiedenen Feldern liegen.

Info: Grenzen vorab festlegen, jederzeit stoppen können

Ein Agent muss nicht bei jedem Schritt beobachtet werden. Die wirksamste Kontrolle findet vor dem Einsatz statt, indem festgelegt wird, welche Rechte, Daten und Werkzeuge er erhält und an welchen Stellen ein Mensch freigeben muss. Unabhängig davon sollte jedes agentische System jederzeit gestoppt werden können. Für Hochrisiko-Systeme verlangt der EU AI Act eine solche Unterbrechungsmöglichkeit ausdrücklich (Art. 14) [23].

Info: Vier Fragen zur Einordnung eines Vorhabens

1. Wer oder was startet das System: ein Mensch, ein Zeitplan oder ein Ereignis?
2. Erhält es einen festen Auftrag, oder entscheidet es selbst, was zu tun ist?
3. Was kann es ohne Rückfrage bewirken, und lässt sich das rückgängig machen?
4. Wer kann das System jederzeit stoppen, und wie schnell?

2 EINSATZFELDER AGENTISCHER KI IN UNTERNEHMEN

Agentische KI ist nicht auf einen bestimmten Unternehmensbereich beschränkt. Sie kann grundsätzlich überall dort eingesetzt werden, wo Informationen aufgenommen, mit weiteren Daten verknüpft, bewertet und anschließend in eine Handlung oder Entscheidungsvorlage überführt werden müssen. Entscheidend ist weniger, ob ein Unternehmen einer bestimmten Branche angehört, sondern ob die jeweilige Aufgabe eine Struktur besitzt, die ein Agent sinnvoll bearbeiten kann. Besonders geeignet sind Aufgaben, die wiederkehrend auftreten, auf vorhandenen Informationen beruhen und deren Zwischenergebnisse fachlich oder sogar automatisiert überprüft werden können. In vielen Fällen ist sie eher ein unterstützendes System, das Informationen zusammensucht, Arbeitsschritte vorbereitet, Entwürfe erstellt oder Prüfprozesse begleitet. Der Mensch gibt dabei weiterhin das Ziel vor, kontrolliert kritische Ergebnisse und bleibt für die fachliche Bewertung verantwortlich. Wie groß der Handlungsspielraum eines Agenten sein darf, hängt vom Prozess, von den verwendeten Daten und von den möglichen Folgen eines Fehlers ab. Ein kurzer und prägnanter Überblick, ob sich eine Aufgabe für den Einsatz von KI eignet, kann Abb. 4 entnommen werden. Entscheidend dabei ist, wie gut eine Aufgabe definiert ist und ob Fehler toleriert werden können. Der letzte Abschnitt dieses Kapitels behandelt zudem noch die rechtlich relevante Einstufung von KI-Anwendungen nach dem AI Act der Europäischen Union (siehe Kap. 2.7 zum EU AI Act KI-Verordnung, Verordnung (EU) 2024/1689).

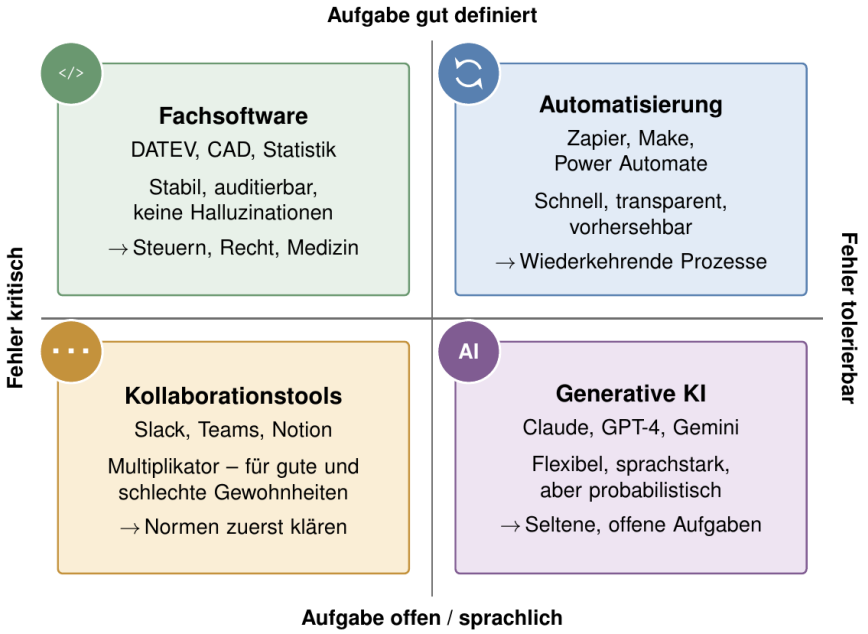

Abbildung 4: Überblick zur Anwendbarkeit von KI anhand von Aufgabendefinition und Fehlertoleranz [24].

Im Folgenden werden mögliche Einsatzbereiche gezeigt, in denen agentische KI in Unternehmen eingesetzt werden kann. Sie sind nicht als vollständige Liste zu verstehen, sondern sollen verdeutlichen, welche Arten von Aufgaben sich für agentische Systeme eignen und auch zum eigenständigen Überlegen zu möglichen Problemen anregen. Gleichzeitig zeigen sie, dass der Einsatz nicht nur nach dem möglichen Nutzen bewertet werden sollte, sondern ebenso einbeziehen, welche Daten verwendet werden, welche Rechte ein Agent erhält, an welchen Stellen menschliche Kontrolle notwendig bleibt und welche Risiken sich aus dem Einsatz ergeben können. Zusätzlich werden auch einige Beispiele eingestreut, die zunächst etwas abstrakt klingende Inhalte nochmal deutlich greifbarer machen sollen und auch auf konkrete Risiken eingehen. Nicht jede der folgenden Aufgaben erfordert zwingend agentische KI. Wenn Datenquelle, Arbeitsschritte und Ergebnisform bereits vollständig feststehen, ist ein klassischer Workflow häufig einfacher, günstiger und zuverlässiger. Ein Agent wird erst dann sinnvoll, wenn das System selbst bestimmen muss, welche Informationen benötigt werden, welche Quellen oder Werkzeuge dafür geeignet sind und wie nach einem Zwischenergebnis weitergearbeitet werden soll.

2.1 INFORMATIONS- UND WISSENSARBEIT

Das naheliegendste Einsatzfeld für KI ist die Verarbeitung, Aufbereitung und Recherche von Informationen. Selten liegen bei Entscheidungsprozessen alle Informationen offen und können ohne weiteres für die richtige Entscheidung verwendet werden. Oft ist eine Vielzahl von technischen Dokumentationen, internen Richtlinien, Berichten, Protokollen, Normen, Verträgen oder regulatorischen Anforderungen zu berücksichtigen. Der Unterschied zu einer einfachen Suchfunktion besteht darin, dass ein Agent Informationen nicht nur findet, sondern weiterverarbeitet. Er kann aus

einer Fragestellung ableiten, welche Informationen benötigt werden, geeignete Quellen durchsuchen, Ergebnisse miteinander vergleichen und daraus eine strukturierte Zusammenfassung erstellen. Aus verstreuten Dokumenten entsteht dadurch nicht automatisch neues Wissen, aber eine besser nutzbare Grundlage für weitere Arbeitsschritte. Der Nutzen liegt deshalb vor allem dort, wo Informationen zwar vorhanden sind, aber über verschiedene Systeme, Dokumente oder Zuständigkeiten verteilt liegen.

Beispiel: Kundenfeedback zusammenführen und auswerten

Ein Unternehmen möchte einen Überblick über das Feedback zur eigenen Organisation aus unterschiedlichsten Quellen gewinnen. Die Kanäle, über die Rückmeldungen eingehen, sind mannigfaltig: Supportanfragen, Händlerberichte, Kundenumfragen, öffentliche Bewertungen, Beiträge in sozialen Medien, Reklamationen oder direkte Gespräche mit dem Vertrieb. Jede dieser Quellen bildet nur einen Ausschnitt der tatsächlichen Wahrnehmung ab. Zusätzlich unterscheiden sich die Rückmeldungen stark in Form, Umfang und Genauigkeit. Eine formale Reklamation enthält häufig konkrete Angaben zu Produkt, Zeitpunkt und Fehlerbild, während eine öffentliche Bewertung oft nur eine kurze subjektive Einschätzung wiedergibt. Zusätzlich ist die Wahrnehmung dieser Rückmeldungen in einer Organisation wahrscheinlich verteilt über Vertrieb, Support, Marketing, Qualitätsmanagement und Geschäftsführung. Alle nehmen unterschiedliche Quellen wahr und gewichten diese häufig vor dem Hintergrund ihrer eigenen Aufgaben. Hier kann ein agentisches KI-System helfen zu vereinheitlichen und den Bias zu minimieren.
Die Zusammenstellung eines Überblicks über Feedback zur eigenen Organisation ist eine für agentische KI gut geeignete Aufgabe. Voraussetzung hierfür ist der Zugriff auf definierte Datenquellen, zum Beispiel Exporte aus CRM- und Supportsystemen, Händlerberichte, Umfrageergebnisse, freigegebene Dokumentenablagen oder zulässig angebundene Social-Media-Kanäle. Eine KI kann hier in einem ersten Schritt alle Quellen nach einem einheitlichen Schema zusammenführen. Das könnte im einfachsten Fall eine Tabelle mit Feldern wie Quelle, Datum, Produkt oder Dienstleistung, betroffener Prozessschritt, erkannter Themenbereich, konkrete Rückmeldung, Bewertung, Häufigkeit, zeitlicher Verlauf und Unsicherheit der Zuordnung sein. Bei den genannten Feldern sind nun auch schon solche enthalten, die die KI eigenständig interpretiert hat. Diese sind vor allem dann wertvoll, wenn man die Analyse noch ein wenig weiter ausführen möchte. Diese Felder können dann zum Filtern nach Themen genutzt werden.

Risiko: Datenbereitstellung und Pflege

Je nachdem wie gut die beteiligten Abteilungen bereits vernetzt sind und in welcher Form die notwendigen Daten bereits vorliegen, kann einiger Aufwand notwendig sein, um ein System bereitzustellen, bei dem die KI Zugriff auf alle wichtigen Quellen hat und die Mitarbeitenden die notwendigen Informationen kontinuierlich ablegen können. Auch wenn die KI durchaus in der Lage ist heterogene Datensätze zu verarbeiten, können auch hier Vereinheitlichungen das Ergebnis auf Kosten zusätzlicher Arbeitszeit verbessern.

Für weitergehende Auswertungen sollten die Rückmeldungen in einer strukturierten Datensammlung mit festgelegten Feldern und Datentypen abgelegt werden. Dafür genügt je nach Umfang bereits ein geeignetes Dateiformat, beispielsweise SQLite. Bei kleinen und einfachen Datensammlungen kann auch eine CSV-Datei ausreichen; eine komplexe Datenbankinfrastruktur ist nicht zwingend erforderlich. Mit dieser Variante kann ein Agent zum Beispiel auch die Häufigkeit von bestimmten Themen oder eine allgemeine Änderung der Außenwahrnehmung in einem zeitlichen Kontext einordnen, falls aus dem gesammelten Feedback nicht direkt hervorgeht, woher diese Änderung rührt. An diesem Punkt sind die Möglichkeiten zur Auswertung vielfältig und die Stärke von KI, Daten zu verarbeiten, kommt zum Tragen.

2.2 KOMMUNIKATION UND KUNDENKONTAKT

Kommunikation mit Kunden ist ein weiteres Feld, in dem agentische KI sinnvoll eingesetzt werden kann. Viele Kundenanfragen bestehen nicht nur aus einer einzelnen Frage, sondern erfordern die Einordnung eines Sachverhalts. Ein einfacher Chatbot kann solche Zusammenhänge nur begrenzt berücksichtigen, wenn er keinen Zugriff auf die relevanten Systeme und Dokumente besitzt.
Ein agentisches System kann eingehende Anfragen zunächst klassifizieren und anschließend die für eine Antwort notwendigen Informationen beschaffen. Es kann zum Beispiel prüfen, ob eine Anfrage technischer, kaufmännischer oder organisatorischer Natur ist, passende Einträge aus Ticketsystem, Produktdokumentation oder Wissensdatenbank hinzuziehen und daraus einen Antwortentwurf erstellen. Zusätzlich kann es offene Punkte markieren, Rückfragen vorschlagen oder den Vorgang an eine zuständige Person weiterleiten. Der Nutzen liegt nicht darin, den Kundenkontakt vollständig zu automatisieren, sondern wiederkehrende Bearbeitungsschritte schneller und einheitlicher vorzubereiten.

Risiko: Kommunikation nach außen und Freigabeprozesse

Für den Einsatz im Kundenkontakt sollten klare Regeln festgelegt werden, welche Antworten automatisch versendet werden dürfen, welche nur als Entwurf erstellt werden und bei welchen Themen zwingend eine menschliche Freigabe erforderlich ist. Besonders bei Beschwerden, Vertragsfragen, Preisen, Lieferzusagen, technischen Zusicherungen oder personenbezogenen Daten sollte die KI nur unterstützend eingesetzt werden. Die Verantwortung für die Kommunikation gegenüber Kunden bleibt grundsätzlich beim Unternehmen. Das System kann Freigabeprozesse unterstützen, indem es Anfragen nach Thema, Risiko oder Zuständigkeit kategorisiert. Fehlerhafte Kategorisierungen sind jedoch nicht auszuschließen. Zusätzlich müssen Tonalität, Verantwortlichkeiten, Protokollierung und Zugriff auf Kundendaten geregelt sein.

Beispiel: Technische Serviceanfrage vorbereiten

Ein Kunde einer Firma für Verpackungsmaschinen meldet per E-Mail, dass eine Maschine wiederholt mit derselben Störmeldung ausfällt und bittet um Unterstützung. Der Agent ordnet die Nachricht als technische Serviceanfrage ein und versucht anhand der vorhandenen Angaben, die betroffene Maschine und ihre Historie im Ticketsystem zuzuordnen. Dabei stellt er beispielsweise fest, dass dieselbe Störung bei Maschinen dieser Baureihe in den vergangenen drei Monaten deutlich häufiger dokumentiert wurde als zuvor. Aus den zugehörigen Wartungs- und Servicetickets erkennt er

außerdem, dass der Anstieg zeitlich mit der Einführung einer neuen Softwareversion zusammenfällt. Der Agent kann diesen Zusammenhang als mögliche Ursache kennzeichnen, passende Abschnitte aus Produkt- und Wartungsdokumentation heranziehen und prüfen, ob für die weitere Eingrenzung noch Informationen fehlen. Ist beispielsweise die aktuell installierte Softwareversion des Kunden nicht bekannt, erstellt er einen Entwurf für eine gezielte Rückfrage. Liegen die notwendigen Angaben bereits vor, bereitet er für das Serviceteam einen strukturierten Vorgang mit Maschinenhistorie, auffälligen Zusammenhängen, verwendeten Quellen und einem Antwortentwurf vor. Ob tatsächlich ein Zusammenhang mit dem Softwareupdate besteht und welche technische Aussage gegenüber dem Kunden getroffen wird, entscheidet weiterhin das Fachpersonal.

2.3 VERWALTUNG UND KOORDINATION

Verwaltung und Koordination gehören zu den Bereichen, in denen agentische KI vor allem durch die Verbindung mehrerer kleiner, wiederkehrender Arbeitsschritte nützlich werden kann. Diese Bereiche sind nicht deshalb aufwendig, weil die einzelnen Aufgaben besonders schwierig wären, sondern weil ständig zusammengehörende Informationen auseinanderfallen. Eine Rechnung liegt in der Buchhaltung, die Bestellung im Einkauf, der Lieferschein im Lager, die Rückfrage dazu in einer E-Mail und die eigentliche Entscheidung wurde mündlich in einem Termin getroffen. Ein Mensch kann solche Zusammenhänge rekonstruieren, aber es ist keine besonders wertschöpfende Arbeit. Ein KI-Agent kann zu einem Vorgang selbstständig die wahrscheinlich zugehörigen Unterlagen suchen, fehlende Bestandteile erkennen und abhängig vom Ergebnis weitere Quellen oder Zuständigkeiten prüfen. In der Buchhaltung und steuerlichen Vorbereitung ist das besonders naheliegend. Die KI sollte nicht die Steuerberatung ersetzen und auch keine endgültige steuerliche Bewertung treffen. Sie kann aber prüfen, ob zu einer Eingangsrechnung eine Bestellung, ein Lieferschein und ein Zahlungsnachweis vorhanden sind, ob Beträge oder Mengen offensichtlich voneinander abweichen und ob Belege für die weitere Bearbeitung vollständig erscheinen. Aus Sicht der Mitarbeitenden verschwindet damit nicht die Verantwortung, aber ein erheblicher Teil der Such- und Sortierarbeit. Neu entsteht dafür die Pflicht, Belege und Vorgänge so abzulegen, dass das System sie überhaupt sinnvoll zuordnen kann. Schlechte Ablage wird durch KI nicht besser, aber schneller sichtbar.

Eine weitgehende Automatisierung von Verwaltung und Koordination ist naheliegend, aber nicht unproblematisch. Die kritische Natur der behandelten Daten, ganz besonders bei Personal-, Finanz- und Vertragsdaten, fordert aber (noch) Zurückhaltung. Stattdessen sollte ein Agent für nüchterne Ordnung sorgen. Er führt verstreute Vorgänge zusammen, markiert fehlende Unterlagen und überträgt Aufgaben in bestehende Systeme. Der Agent übernimmt in diesen Fällen nicht die Verantwortung für den Prozess, sondern reduziert den Aufwand, der durch Suchen, Erinnern, Zuordnen und Zusammenfassen entsteht.

Beispiel: Nachweise und Fristen überwachen

Ein mittelständischer Produktionsbetrieb muss für Mitarbeitende und Betriebsmittel zahlreiche wiederkehrende Nachweise aktuell halten. Dazu gehören beispielsweise Sicherheitsunterweisungen, Qualifikationsnachweise für bestimmte Tätigkeiten, Prüfungen von Arbeitsmitteln oder kundenseitig

geforderte Zertifikate. Die Informationen liegen in Personalakten, Schulungsunterlagen, Prüfprotokollen und verschiedenen Ablagen. Ein Agent kann diese Quellen regelmäßig auswerten, Gültigkeitszeiträume erfassen und prüfen, welche Nachweise für welche Personen, Tätigkeiten oder Betriebsmittel erforderlich sind. Er erkennt beispielsweise, dass die Qualifikation eines Mitarbeiters in sechs Wochen ausläuft, dieser aber bereits für einen späteren Auftrag eingeplant ist, bei dem genau dieser Nachweis benötigt wird. Statt lediglich einen Kalendereintrag zu erzeugen, kann der Agent den gesamten Vorgang vorbereiten. Er sucht den letzten Nachweis, ermittelt die zuständige Stelle, prüft verfügbare Termine oder interne Vorgaben und erstellt eine Aufgabe oder einen Entwurf zur Verlängerung. Fehlt ein Nachweis vollständig oder lässt sich eine Anforderung nicht eindeutig zuordnen, wird der Fall zur Prüfung vorgelegt. Nach erfolgter Schulung oder Prüfung ordnet der Agent das neue Dokument dem bestehenden Vorgang zu und aktualisiert den überwachten Friststatus. Der Verwaltungsaufwand verschiebt sich dadurch von der fortlaufenden Suche und Fristenkontrolle auf die Bearbeitung der tatsächlich auffälligen oder entscheidungsbedürftigen Fälle.

2.4 EINKAUF UND LIEFERANTENMANAGEMENT

Der Einsatz agentischer KI im Einkauf ist vor allem dort interessant, wo Mitarbeitende einen großen Teil ihrer Arbeit mit dem Suchen, Zuordnen und Vergleichen von Informationen verbringen. Bei einer Anfrage müssen häufig frühere Angebote, bestehende Lieferanten, alte Bestellungen, technische Anforderungen, Lieferzeiten, Reklamationen oder offene Rückfragen berücksichtigt werden. Diese Informationen müssen aber erst aus verschiedenen Vorgängen und Systemen zusammengesucht werden. Ein Agent kann abhängig von der jeweiligen Anfrage selbst bestimmen, welche früheren Vorgänge, Angebote und Lieferantenerfahrungen relevant sind, diese zusammenführen und fehlende Informationen gezielt nachfordern. Ein Agent kann zum Beispiel zeigen, dass ein Anbieter in früheren Projekten zwar günstige Angebote abgegeben hat, aber häufiger verspätet geliefert oder zusätzliche Abstimmungen verursacht hat. Die Bewertung, ob dieser Anbieter trotzdem geeignet ist, bleibt eine kaufmännische und fachliche Entscheidung. Der Nutzen der KI liegt darin, die Grundlage der Auswahl zu verbessern. Entscheidungen sind weniger von Einzelpreisen, persönlicher Erinnerung oder unvollständigen Informationen abhängig und bei eventueller Vertretung oder Personalwechsel werden die Prozesse stark unterstützt.

Der Einsatz von agentischer KI verändert jedoch auch die Arbeit der Mitarbeitenden. Such- und Zusammenstellungsarbeit kann reduziert werden, dafür entstehen neue Aufgaben bei der Pflege der Datenbasis und bei der Prüfung der KI-Ergebnisse. Frühere Vorgänge müssen so abgelegt sein, dass ein Agent sie zuordnen kann. Lieferantenbewertungen, Reklamationen, Nacharbeiten oder Lieferverzögerungen müssen dokumentiert werden, wenn sie später in Entscheidungen einfließen sollen. Sobald KI in diese Prozesse eingebunden ist, kann sie natürlich auch bei der Dokumentation helfen. Mitarbeitende müssen statt dieser Aufgaben prüfen, ob die vom System vorgeschlagenen Zusammenhänge stimmen und ob wichtige Informationen fehlen. Agentische KI verschiebt Arbeit daher teilweise von der Informationssuche zur Informationspflege und Bewertung.

Für kleine Unternehmen mit wenigen Lieferanten und überschaubaren Abläufen kann der Einsatz von KI an dieser Stelle schnell überdimensioniert sein. Für KMU mit vielen wiederkehrenden Dienstleistern, projektbezogenen Einkäufen oder mehreren Standorten kann es dagegen sinnvoll werden, besonders wenn Einkaufsentscheidungen bisher stark über Einzelpreise getroffen werden, während Folgekosten kaum systematisch erfasst werden. Die KI kann dann helfen, nicht nur den günstigsten Einkaufspreis, sondern auch spätere Kosten durch Verzögerungen, Nacharbeiten oder zusätzlichen Abstimmungsaufwand sichtbarer zu machen.

Beispiel: Preiserhöhung des Produktionsmaterials

Ein wichtiger Lieferant kündigt kurzfristig eine deutliche Preiserhöhung für ein regelmäßig benötigtes Produktionsmaterial an. Weil der Agent die zugehörige Lieferantenakte bereits fortlaufend aus Bestellungen, Lieferzeiten, Reklamationen, Preisentwicklungen und früheren Abstimmungen ergänzt hat, muss der Einkauf diese Informationen nicht erst nach Eingang der Mitteilung zusammensuchen. Der Agent kann nun zusätzlich angewiesen werden, auf Basis des bisherigen Verbrauchs, geplanter Bedarfe und des neuen Preises abzuschätzen, welche Mehrkosten in den kommenden Monaten entstehen würden. Gleichzeitig lässt sich erkennen, wie zuverlässig der Lieferant bisher gearbeitet hat und welche alternativen Anbieter oder früheren Vergleichsangebote bekannt sind. Gerade wenn die eigene Produktion auf dieses Material angewiesen ist, entsteht dadurch ein erheblicher Zeitgewinn. Der Agent kann verschiedene Optionen gegenüberstellen, etwa die Preiserhöhung zu akzeptieren, nachzuverhandeln oder einen Lieferantenwechsel vorzubereiten, und daraus eine Entscheidungsvorlage erstellen. Die eigentliche kaufmännische und strategische Entscheidung bleibt beim Einkauf. Der Nutzen des Agenten zeigt sich hier weniger in einer einzelnen Analyse als darin, dass über längere Zeit nebenbei eine belastbare Informationsbasis gepflegt wird, die bei einem plötzlich auftretenden Problem unmittelbar genutzt werden kann.

2.5 IT-BETRIEB UND SUPPORT

IT-Betrieb besteht zu einem großen Teil aus der Auswertung bereits vorhandener Informationen. Monitoring-Systeme melden Auffälligkeiten, Firewalls und Server schreiben Protokolle, Geräte besitzen unterschiedliche Firmwarestände und Konfigurationen werden über Jahre verändert. Ein Agent kann diese Quellen gemeinsam betrachten und beispielsweise feststellen, dass wiederkehrende Verbindungsabbrüche erst seit einer Änderung an einem Switch auftreten oder dass eine Firewall-Regel nicht mehr zur dokumentierten Netzstruktur passt. Für geplante Änderungen kann der Agent die betroffenen Systeme bestimmen, Konfigurationen vergleichen und einen Ablauf mit Prüfschritten und Rückfallmöglichkeit vorbereiten. Eine produktive Netzwerkkonfiguration sollte er deshalb noch lange nicht selbstständig umschreiben.

Nutzer-Support verhält sich ähnlich wie IT-Betrieb, aber in kleinerem Maßstab. Ein Agent ergänzt unvollständige Tickets, sucht nach vergleichbaren Fällen und stellt dem IT-Personal die wahrscheinlich relevanten Informationen bereit. Dieselbe Arbeitsweise lässt sich auf Geräteinventar, offene Updates, Zugriffsrechte oder fehlgeschlagene Backups übertragen. Der Agent übernimmt vor allem die fortlaufende Sichtung, während ungewöhnliche Fehler, Sicherheitsentscheidungen und Änderungen

mit größerer Reichweite beim Fachpersonal bleiben. Die Einführung sollte mit lesendem Zugriff und einem eigenen, stark begrenzten Dienstkonto beginnen. Erst nach ausreichend vielen überprüften Ergebnissen sind reversible Aktionen sinnvoll, etwa ein Diagnoseprogramm zu starten, ein Ticket anzulegen oder eine freigegebene Standardprozedur auszuführen. Vollständige Administratorrechte würden zwar jede Automatisierung vereinfachen, aber auch aus einem einzelnen Modellfehler einen unternehmensweiten Ausfall machen. Der wirtschaftliche Nutzen hängt deshalb weniger von möglichst großer Autonomie ab als von einer sauberen Abstufung der Rechte: häufige und rückgängig zu machende Arbeiten automatisieren, seltene und folgenreiche Eingriffe prüfen.

Beispiel: Migration eines internen Dienstes

Ein Unternehmen möchte einen intern betriebenen Dienst auf einen neuen virtuellen Server umziehen. Der Agent ermittelt zunächst aus der vorhandenen Dokumentation und den Systemkonfigurationen, welche Abhängigkeiten bestehen, und richtet anschließend innerhalb der freigegebenen Umgebung einen neuen Server ein. Er installiert die benötigte Software, überträgt Konfigurationen und zunächst eine Kopie der Daten und führt definierte Funktionstests durch. Treten dabei Fehler auf, kann er Protokolle auswerten, Konfigurationen anpassen und die Tests erneut ausführen. Aus den Ergebnissen erstellt er schließlich einen konkreten Ablauf für die eigentliche Umstellung einschließlich Prüfschritten und Rückfallmöglichkeit.
Die produktive Umschaltung erfolgt erst nach Freigabe durch den Administrator. Anschließend kann der Agent beispielsweise die letzte Datensynchronisation durchführen, den Dienst auf dem neuen System starten, dessen Erreichbarkeit und wesentliche Funktionen prüfen und die tatsächlich ausgeführten Änderungen dokumentieren. Ein erheblicher Teil der technischen Arbeit wird damit nicht nur vorbereitet, sondern durch den Agenten selbst über Administrationswerkzeuge, Skripte und Kommandozeilen ausgeführt. Der Administrator konzentriert sich auf die Bewertung des Migrationsplans, kritische Freigaben und ungewöhnliche Fehler, während wiederholbare Installations-, Konfigurations-, Prüf- und Dokumentationsschritte weitgehend vom Agenten übernommen werden.

2.6 EINORDNUNG DER BEISPIELE

Die in diesem Kapitel vorgestellten Beispiele unterscheiden sich stark anhand des Integrationsaufwands und der zu beachtenden rechtlichen Bestimmungen. Beide Merkmale sind kritisch für eine Abschätzung, ob sich der Aufwand am Ende lohnt. Der technische Integrationsaufwand steigt insbesondere dann, wenn Informationen aus vielen unterschiedlichen Systemen zusammengeführt werden müssen, geeignete Schnittstellen fehlen oder der Agent nicht nur lesen, sondern selbst Änderungen ausführen soll. Eine bestehende, gut strukturierte Datenbasis kann den Aufwand erheblich reduzieren. Auch die Stabilität des Prozesses ist wichtig: Häufig wechselnde Datenquellen, Regeln oder Software erhöhen den späteren Wartungsaufwand. Für rechtliche Anforderungen sind beispielsweise personenbezogene oder vertrauliche Daten, gesetzliche Nachweispflichten, wirtschaftlich bindende Entscheidungen und Eingriffe in produktive Systeme relevant. Besonders wichtig ist zudem die Reversibilität einer Handlung. Ein fehlerhafter interner Bericht lässt sich vergleichsweise einfach korrigieren, während eine falsche Änderung an einem

produktiven IT-System unmittelbar einen Ausfall verursachen kann. Menschliche Freigaben können Risiken deutlich begrenzen, beseitigen sie jedoch nicht vollständig.

In Abb. 5 werden die Beispiele anhand des Integrationsaufwands und der zu beachtenden rechtlichen Bestimmungen qualitativ eingeordnet. Die Überwachung von Nachweisen und Fristen benötigt vergleichsweise wenig technische Integration, kann wegen gesetzlicher oder betrieblicher Pflichten aber dennoch kritisch sein. Eine Zusammenführung von Kundenfeedback ist rechtlich und betrieblich meist weniger riskant, erfordert bei vielen unterschiedlichen Datenquellen jedoch einen erheblichen Integrationsaufwand. Die Migration eines internen Dienstes liegt bei beiden Merkmalen deutlich höher, weil der Agent nicht nur Informationen verarbeitet, sondern selbst Konfigurationen ändern, Daten übertragen und Systemzustände beeinflussen kann. Für einen ersten Einsatz eignen sich daher besonders Prozesse, bei denen der Agent zunächst lesend arbeitet, Ergebnisse gut überprüfbar sind und Fehler nur begrenzte Folgen haben. Ein technisch einfacher Prozess ist dagegen nicht automatisch ein guter Einstieg, wenn bereits kleine Fehler rechtliche oder betriebliche Konsequenzen auslösen können.

Nach der Einordnung eines Prozesses kann schlussendlich die Überlegung, ob sich der Prozess auch wirtschaftlich umsetzen lässt, vorgenommen werden. Wirtschaftlich interessant sind vor allem Prozesse, die häufig auftreten oder pro Vorgang einen hohen manuellen Aufwand verursachen. Ein hoher Nutzen kann auch bei seltenen Vorgängen entstehen, wenn Verzögerungen oder Fehler erhebliche Folgekosten verursachen. Besonders günstig sind Anwendungen, bei denen einmal geschaffene Schnittstellen und Datenzugriffe für mehrere Prozesse wiederverwendet werden können und die Ergebnisse mit geringem Aufwand kontrollierbar bleiben. Ein Agent muss dabei nicht jeden Sonderfall vollständig beherrschen. Bereits die zuverlässige Bearbeitung oder Vorbereitung eines großen Anteils der wiederkehrenden Fälle kann einen erheblichen wirtschaftlichen Nutzen erzeugen.

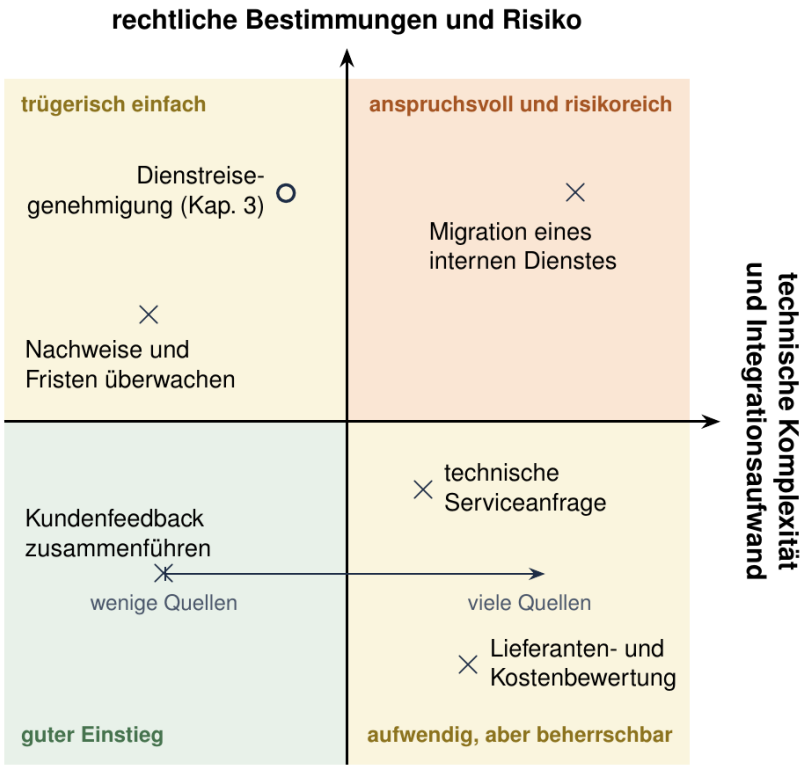

Abbildung 5: Einordnung der Beispiele hinsichtlich technischer Komplexität bzw. Integrationsaufwand und rechtlicher Bestimmungen bzw. Risiko. Der Pfeil beim Kundenfeedback zeigt, wie der Aufwand mit der Zahl der angebundenen Quellen wächst. Das Beispiel aus Kap. 3 ist mit einem Kreis markiert.

2.7 DER EU AI ACT

Im Jahr 2024 trat mit dem EU AI Act die erste umfassende, EU-weit geltende Regelung zur Nutzung und Bereitstellung von KI-Systemen in Kraft. Die Verordnung knüpft Anforderungen insbesondere an den vorgesehenen Verwendungszweck eines Systems, um das Risiko des Einsatzes abzuschätzen. Für eine erste Einordnung sind insbesondere Artikel 5 zu verbotenen Praktiken, Artikel 6 zur Einstufung als Hochrisiko-KI sowie Anhang III relevant. Die Europäische Kommission stellt darüber hinaus Leitlinien und konkrete Anwendungsbeispiele zur Verfügung, die bei der Einordnung unterstützen sollen. Als Hochrisiko gelten KI-Systeme, die selbst ein bestimmtes reguliertes Produkt oder dessen Sicherheitskomponente darstellen und dadurch auch einer vorgeschriebenen Drittstellenprüfung unterliegen. Anhang III nennt konkrete Einsatzgebiete, darunter bestimmte Anwendungen in kritischer Infrastruktur, Bildung sowie Beschäftigung und Personalmanagement. Eine Ausnahme von der Einstufung als Hochrisiko-System kommt für die in Artikel 6 Absatz 3 genannten Fallgruppen nur in Betracht, wenn das System kein erhebliches Schadensrisiko für die Gesundheit, Sicherheit oder Grundrechte natürlicher Personen begründet und insbesondere das Ergebnis einer Entscheidung nicht wesentlich beeinflusst. Zu diesen Fallgruppen gehören unter anderem eng begrenzte Verfahrensaufgaben und vorbereitende Tätigkeiten. Führt das System innerhalb eines Anwendungsfalls aus Anhang III ein Profiling natürlicher Personen durch, bleibt die Ausnahme

ausgeschlossen. Ein Anbieter, der ein in Anhang III genanntes System dennoch nicht als hochriskant einstuft, muss diese Bewertung vor dem Inverkehrbringen oder der Inbetriebnahme dokumentieren und auf Verlangen der zuständigen Behörde vorlegen [23].
Die Einstufung hat konkrete Folgen für den Betrieb. Vor der Einführung sind jedoch auch die Rollen nach dem EU AI Act zu bestimmen. Ein Unternehmen, das ein KI-System selbst entwickelt oder entwickeln lässt und unter eigenem Namen in Betrieb nimmt, kann auch dann als Anbieter des Gesamtsystems gelten, wenn ein extern bereitgestelltes Sprachmodell verwendet wird oder das System nur intern eingesetzt wird; bei Hochrisiko-Systemen können dadurch zusätzlich zu den Betreiberpflichten auch Anbieterpflichten entstehen. Die zentralen Anforderungen an Hochrisiko-Systeme aus Kapitel III gelten nach der Verschiebung durch den Digital Omnibus für Systeme nach Artikel 6 Absatz 2 und Anhang III ab dem 2. Dezember 2027. Für Hochrisiko-Systeme nach Artikel 6 Absatz 1 und Anhang I, insbesondere KI-Systeme in bestimmten regulierten Produkten oder deren Sicherheitskomponenten, gelten sie ab dem 2. August 2028. Für Betreiber von Hochrisiko-Systemen sieht Artikel 26 unter anderem geeignete technische und organisatorische Maßnahmen, qualifizierte menschliche Aufsicht, Überwachung des Systems und die Aufbewahrung bestimmter Protokolle vor. Bei erkannten erheblichen Risiken oder schwerwiegenden Vorfällen bestehen zusätzlich Informations- beziehungsweise Meldepflichten. Unabhängig von einer Hochrisiko-Einstufung können Systeme besonderen Transparenzpflichten unterliegen. Seit August 2026 verlangt Artikel 50 in bestimmten Fällen, dass Menschen darüber informiert werden, wenn sie unmittelbar mit einem KI-System interagieren oder mit bestimmten KI-generierten beziehungsweise manipulierten Inhalten konfrontiert werden [23]. Unabhängig von der Risikoklasse gilt bereits seit dem 2. Februar 2025 Artikel 4. Anbieter und Betreiber müssen Maßnahmen zur Förderung einer für den jeweiligen Einsatz angemessenen KI-Kompetenz der mit den Systemen arbeitenden Personen ergreifen.

3 EINFÜHRUNG EINES AGENTISCHEN KI-SYSTEMS

Auch wenn das letzte Kapitel schon viel über verschiedene Einsatzmöglichkeiten von KI gesprochen hat, soll dieses Kapitel nochmal ein konkretes Beispiel aufgreifen und anhand dessen die Einführung/Umsetzung von agentischer KI zeigen. Das hier gewählte Beispiel soll der Prozess einer Dienstreisegenehmigung und der anschließenden Abrechnung sein. Das Ziel des Einsatzes eines Agenten soll sein, den Prozess in großen Schritten zu automatisieren, aber finale Freigaben und Kommunikation weiterhin dem Personal zu überlassen. Dies ist in Abb. 6 illustriert. Die eingehenden Vorgänge sollen automatisch einen Hintergrundprozess des Agenten starten, der dann den Vorgang autonom so weit treibt, dass ein Mitarbeiter direkt prüfen kann. Dabei greift der Agent auf eine Wissensbasis zurück, auf die er hauptsächlich Lesezugriff hat und die von Fachpersonal gepflegt wird. Abgeschlossene Vorgänge kann der Agent aber ebenfalls eigenständig ablegen. Das Sprachmodell läuft beim Anbieter in der Cloud, der zugehörige Harness wird als Command-Line-Interface lokal installiert und erhält darüber Zugriff auf die relevanten Systeme des Unternehmens erhält und auch

darüber betrieben wird. Dem stehen beispielsweise lokale Modelle entgegen, die nochmal in Kap. 4 aufgegriffen werden.

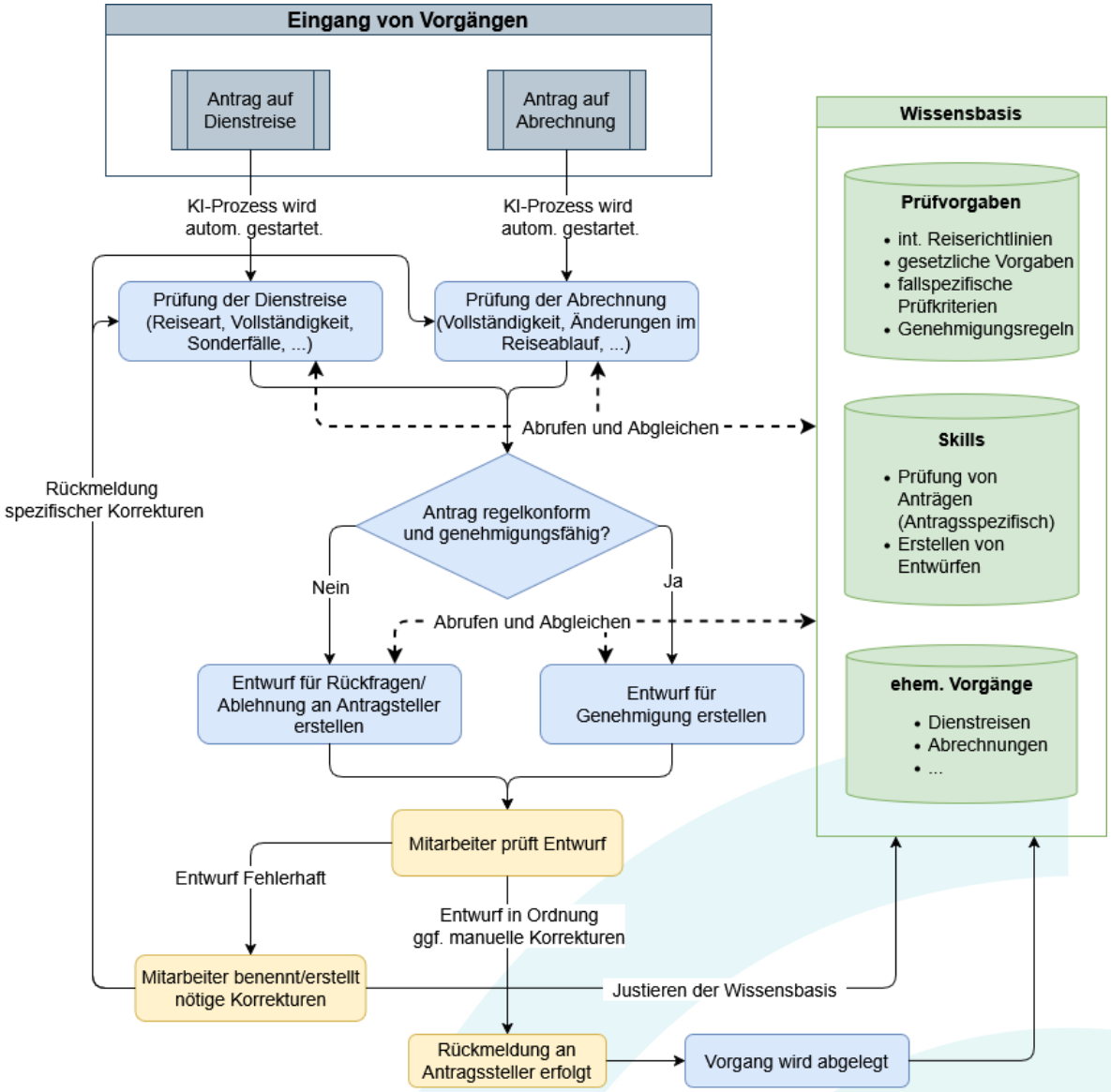

Abbildung 6: Zielbild des Arbeitsprozesses mit agentischer KI als etabliertes Werkzeug (blau) und Kontrolle durch Mitarbeiter (gelb). Der Agent erstellt proaktiv Entwürfe und legt diese dem Personal zur Kontrolle vor.

3.1 IMPLEMENTATION DER AGENTISCHEN KI

Ausgehend von einem üblichen Computer von einem Mitarbeiter wird angenommen, dass es einen lokalen Zugriff auf die Dateiablage für Vorgänge, z.B. über Netzlaufwerk, NAS oder lokale Synchronisation, gibt. Ausgehend hiervon wird zunächst ein *Command-Line-Interface* (CLI) eines passenden Agenten installiert (siehe Kap. 1.3). Ein CLI bildet die Schnittstelle zwischen dem Sprachmodell, das weiterhin auf den Servern des Anbieters läuft, und dem Computer des Mitarbeiters.

Das CLI ist der sogenannte Harness: Es führt die Agentenschleife aus, setzt die Regeln durch, welche Befehle erlaubt sind, und reicht die Ergebnisse an das Sprachmodell zurück. Das Sprachmodell kann weiterhin ausschließlich Text generieren. Um Befehle auf dem Computer auszuführen, muss er diese an das CLI senden. Ein CLI kann auf zwei Arten genutzt werden. Entweder Mitarbeitende starten es interaktiv, um Prüfergebnisse zu betrachten oder konkrete Korrekturen zu veranlassen, oder dieselbe Installation wird ohne Benutzereingabe im sogenannten Headless-Modus aufgerufen. Hierfür startet ein Skript eine neue Agenteninstanz, übergibt ihr feste Arbeitsanweisungen und legt das zu bearbeitende Verzeichnis fest.

Im nächsten Schritt muss der Agent durch einen eingehenden Vorgang automatisch gestartet werden. Wenn Vorgänge über den Mailverkehr eingehen, wird auf dem Rechner ein kleines Triggerprogramm eingerichtet, beispielsweise ein Python-Skript, das als Windows-Dienst oder über die Aufgabenplanung regelmäßig ausgeführt wird. Das Triggerprogramm überwacht das Funktionspostfach über IMAP oder die vom Mailanbieter bereitgestellte Programmierschnittstelle (API). Das Programm fragt ab, ob seit dem letzten Durchlauf eine neue Mail oder ein neues Ticket eingegangen ist. Wird ein neuer Vorgang gefunden, lädt das Triggerprogramm die Nachricht, das Ticket und die zugehörigen Anhänge herunter, erzeugt am vorgesehenen Ort ein neues Vorgangsverzeichnis und startet anschließend den Agenten. Bei einem Ticketsystem ist für eine zuverlässige Integration ebenfalls eine API erforderlich. Die API sollte mindestens neue Tickets auflisten oder filtern sowie Ticketfelder, Kommentare und Anhänge auslesen können. Soll der Agent Entwürfe direkt im System ablegen, werden zusätzlich Schreibfunktionen für interne Kommentare, Anhänge oder eigene Felder benötigt. Bei verbreiteten Ticketsystemen muss diese API normalerweise nicht selbst entwickelt werden. Jira Cloud stellt beispielsweise REST-Schnittstellen, Webhooks und fertige Beispiele bereit. Zendesk bietet ebenfalls REST-APIs, Webhooks, Authentifizierung und Beispielaufrufe. Entwickelt werden muss nur ein kleiner Adapter, der die Felder des konkreten Ticketsystems in die für den Agenten benötigte Vorgangsstruktur überführt.

3.2 WAS DER AGENT WISSEN MUSS - DIE WISSENSBASIS

Um KI-Agenten effektiv Anweisungen zu geben, lohnt es sich eine Wissensbasis aufzubauen, die ihm bei Bedarf wiederkehrende Anweisungen, Lösungsstrategien, zusätzliche Informationen und Richtlinien etc. zur Verfügung stellt. Die unattraktive Alternative bestünde darin dem Agenten wichtige Informationen stets per Prompt mitzuteilen. Das Ziel dieser Wissensbasis ist es daher sowohl die Arbeit mit dem Agenten schlank zu halten, als auch einen stetigen Hebel für die Qualität der Arbeit des Agenten zu haben. Damit ist auch klar, dass die Erstellung und Pflege dieser Wissensbasis eine der wesentlichen neuen Aufgaben der Mitarbeiter sein wird, die den entsprechenden Prozess vorher durchgeführt haben.

Ein Agent kann Informationen über wiederkehrende Arbeitsabläufe über sogenannte Skills erhalten. Ein Skill ist ein Paket aus Arbeitsanweisungen und optional zusätzlichen Dateien. Das hier verwendete Beispiel orientiert sich am Agent-Skills-Format. Je nach aktueller Aufgabe des Agenten kann er eigenständig entscheiden, einen Skill zu laden, falls er ihn für relevant hält, oder er wird explizit

angewiesen, diesen zu nutzen. Der Agent hat stets Informationen über Namen und Beschreibung seiner Skills, um dies zu entscheiden, jedoch keine vollständige Kenntnis über den Inhalt. Bei Agentenumgebungen, die dieses Format unterstützen, besteht ein Skill mindestens aus einem eigenen Verzeichnis und einer darin enthaltenen Datei SKILL.md. Zusätzlich können Referenzdokumente, Skripte und Vorlagen abgelegt werden (Abb. 7). Welche Dateiformate daraus gelesen und welche Skripte ausgeführt werden können, hängt von der jeweiligen Laufzeitumgebung, den verfügbaren Werkzeugen und den vergebenen Berechtigungen ab. Beispielsweise können gesetzliche Vorgaben, interne Reiserichtlinien oder auch weitere Programme, im späteren auch Werkzeuge (Kap. 3.3) genannt, abgelegt werden.

Info: .md-Dateien

Eine Markdown-Datei (.md) ist eine Textdatei, die einige Regeln benutzt, um Texte auf sehr einfache Weise zu formatieren. Das Format ist weit verbreitet und wird von vielen Anwendungen auch automatisch erkannt (siehe markdownguide.org).

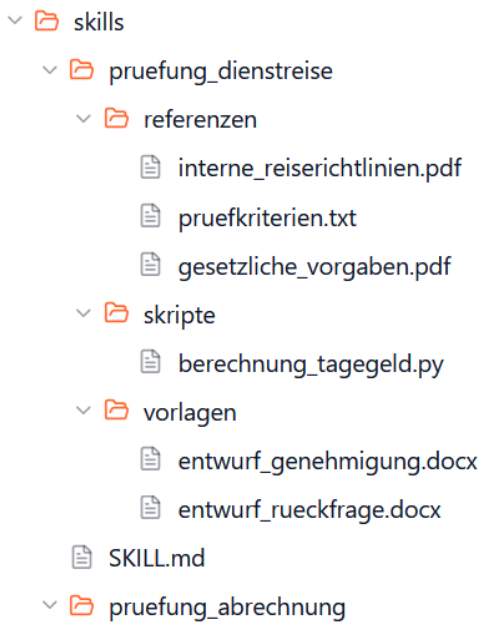

Abbildung 7: Beispiel einer Ordnerstruktur für verschiedene Skills. Pro Skill wird mindestens ein Verzeichnis mit einer `SKILL.md` benötigt.

In der SKILL.md selbst werden zunächst Name und Beschreibung des Skills mit einer spezifischen Syntax hinterlegt und anschließend die verschiedenen Arbeitsanweisungen aufgeführt. Die Idee hinter der Datei ist mehr, die Logik und Reihenfolge festzulegen, als die eigentlich zu bearbeitenden Punkte. Die SKILL.md beschreibt dabei die beabsichtigte Vorgehensweise, erzwingt deren Einhaltung aber nicht technisch. Verbindliche Prüfungen, Freigaben oder Berechtigungsgrenzen sollten daher zusätzlich durch die Laufzeitumgebung oder deterministische Prüfprogramme durchgesetzt werden. Was und wie genau geprüft werden soll, kann in weitere Dateien ausgelagert und hier mit Referenzen,

auch außerhalb des Verzeichnisses, genannt werden, auf die der Agent zugreifen soll. Im Folgenden ein Beispiel, wie eine solche SKILL.md aufgebaut sein kann.

```
SKILL.md

---
name: pruefung_dienstreise
description: Steuert die Prüfung von Dienstreiseanträgen. Verwende diesen Skill
bei neu eingegangenen, ergänzten oder erneut zu prüfenden Dienstreiseanträgen.
---

# Dienstreiseprüfung

## 1. Prüfauftrag vorbereiten

Lies zuerst:
- `referenzen/pruefkriterien.txt`
- `referenzen/datenquellen.md`

Ermittle daraus:
- [...]

Halte das Ergebnis als `Vorgangsdaten` fest.

## 2. Vorgang klassifizieren

[...]

## 3. Vollständigkeit prüfen

Prüfe `Vorgangsdaten` gegen die Anforderungen aus:
`referenzen/pruefkriterien.txt`

Verwende für eindeutig strukturierte Pflichtangaben nach Möglichkeit:
`skripte/pruefung_pflichtangaben.py`

Wenn notwendige Informationen fehlen, springe zu **7. Entwurf erstellen**.

.
.
.
```

3.3 KI WERKZEUGE

Damit ein Agent in seiner Umgebung (dem System, dem Internet, etc.) auch tatsächlich Handlungen ausführen kann, benötigt er Werkzeuge. Dies sind eigenständige, aufrufbare Funktionen, beispielsweise eine Websuche, ein selbst geschriebenes Skript, oder eine Datenbankabfrage. Der Agent entscheidet, wann es aufgerufen wird, das Werkzeug führt eine Aufgabe aus und gibt ein Ergebnis zurück. Der Agent muss danach wieder entscheiden, was er mit dem Ergebnis macht. Im Kontext von KI-Agenten können Werkzeuge also Teile einer Aufgabe aus dem probabilistischen

Sprachmodell in definierte, überprüfbare Verfahren auslagern. Wie stark dadurch die Abhängigkeit vom Agenten sinkt, hängt davon ab, wie viel Entscheidungs- und Verarbeitungslogik bereits im Werkzeug festgelegt ist.

Werkzeuge können lokal oder extern betrieben sowie selbst entwickelt oder von Dritten bereitgestellt werden. Ihre Anbindung an die Agentenanwendung kann beispielsweise über direkte Funktionsaufrufe, APIs oder das Model Context Protocol (MCP) erfolgen. MCP ist dabei kein eigener Bereitstellungsort für Werkzeuge, sondern standardisiert die Kommunikation zwischen einem Client der Agentenanwendung und lokalen oder entfernten Servern, die unter anderem Werkzeuge bereitstellen. MCP ist vor allem dann interessant, wenn eine größere Zahl unterschiedlicher Werkzeuge oder Systeme an einen KI-Agenten angebunden werden sollen. Statt für jedes Werkzeug eine eigene, proprietäre Schnittstelle zur KI zu entwickeln, legt MCP ein gemeinsames Verfahren fest, über das Werkzeuge beschrieben und aufgerufen werden können. Für ein einzelnes, selbst entwickeltes Werkzeug ist MCP daher nicht zwingend notwendig; bei mehreren Systemen oder bereits verfügbaren MCP-Anbindungen kann es den Integrationsaufwand deutlich reduzieren.

Eigene Werkzeuge sind sinnvoll, wenn ein Agent auf unternehmensspezifische Funktionen zugreifen oder bestimmte Aufgaben zuverlässig nach festen Regeln ausführen soll. Dafür kann beispielsweise ein vorhandenes Python-Skript, eine interne API oder eine Datenbankabfrage als Werkzeug bereitgestellt werden. Dem Agenten wird beschrieben, wofür das Werkzeug gedacht ist, welche Eingaben benötigt werden und welches Ergebnis zurückgegeben wird. Ein zusätzliches Training des Sprachmodells ist dafür normalerweise nicht erforderlich. Besonders nützlich ist dies für Aufgaben, deren Ergebnis möglichst wenig vom Sprachmodell abhängen soll. Im Beispiel der Dienstreise könnte der Agent etwa Reiseziel und -dauer bestimmen und anschließend ein fest programmiertes Werkzeug zur Berechnung des Tagegeldes aufrufen. Die Auswahl und Befüllung des Werkzeugs bleibt Aufgabe des Agenten, die eigentliche Berechnung folgt dagegen einem eindeutig definierten Verfahren.

Damit ein Agent die Werkzeuge schlussendlich nutzen kann, wird ihm, zusammen mit den Prompts, die Informationen über alle zur Verfügung stehenden Werkzeuge vom Harness übergeben. Daraus ergibt sich aber auch sofort, dass eine große Auswahl an Werkzeugen nicht automatisch vorteilhaft ist. Müssen dem Agenten viele Werkzeugbeschreibungen gleichzeitig bereitgestellt werden, steigt der Kontextumfang und damit potenziell auch Rechenaufwand und Kosten. Gleichzeitig wird die Auswahl des passenden Werkzeugs schwieriger. Sinnvoll ist daher, die Auswahl gut zu kuratieren. Informationen für lokale Werkzeuge werden über eine Anweisungsdatei, eine Funktion des Harness, die der Agent bei jedem Start als Kontext bekommt, bereitgestellt. Wo diese genau ist, hängt vom jeweiligen Agenten und der Installation ab. Auch wenn in dem Beispiel hier die Berechnung des Tagegeldes mehr Sinn innerhalb des Skill ergibt, wird hier die Information für dieses einfache Python-Skript als Werkzeug gezeigt.

```
Start-Anweisungen (vereinfacht)

## Verfügbare lokale Werkzeuge
```

```
### Tagegeldberechnung

Werkzeug:
`C:\AgentTools\berechnung_tagegeld.py`

Zweck:
Berechnet das Tagegeld einer Dienstreise nach dem hinterlegten Regelwerk. Das Beispiel zeigt eine vereinfachte Schnittstelle; die tatsächlich benötigten Angaben richten sich nach der jeweils umgesetzten Reiserichtlinie beziehungsweise gesetzlichen Grundlage.

Aufruf:
`python C:\AgentTools\berechnung_tagegeld.py --ziel <Ort/Land> --beginn <Datum/Uhrzeit> --ende <Datum/Uhrzeit> --mahlzeiten <Angaben> --regelwerk <Version>`

Eingaben:
- ziel: Zielort beziehungsweise Zielland der Reise
- beginn: Beginn der dienstlichen Abwesenheit
- ende: Ende der dienstlichen Abwesenheit
- mahlzeiten: während der Reise bereitgestellte Mahlzeiten
- regelwerk: anzuwendende Reiserichtlinie beziehungsweise Regelwerksversion

Ausgabe:
JSON mit berechnetem Tagegeld und verwendeten Berechnungsgrundlagen. Fehlen für das hinterlegte Regelwerk notwendige Angaben, wird keine Berechnung durchgeführt, sondern der fehlende Informationsbedarf zurückgegeben.
.
.
.
```

3.4 ZU GUTER LETZT: PROMPTEN

Die Qualität der Ergebnisse eines KI-Systems hängt nicht nur vom verwendeten Modell, sondern auch davon ab, wie eine Aufgabe beschrieben wird. Für einen (teil-)automatisierten Betrieb eines Agenten ist der letzte Schritt die iterative Erstellung aller Arbeitsanweisungen. Komplexe Aufgaben sollten nicht zwangsläufig mit einer einzigen umfangreichen Anweisung gelöst werden. Häufig ist es sinnvoller, die Aufgabe in nachvollziehbare Arbeitsschritte zu zerlegen und Zwischenergebnisse zu prüfen. Ein Ergebnis kann genutzt werden, um unklare Anforderungen zu erkennen und den Arbeitsauftrag anschließend gezielt zu verbessern. Für wiederkehrende Unternehmensprozesse können auf diese Weise erprobte Anweisungen entstehen, die nicht bei jedem Vorgang neu formuliert werden müssen. Vor der produktiven Freigabe sollte das Gesamtsystem jedoch zusätzlich mit getrennten, repräsentativen Testfällen geprüft werden, die nicht bereits zur Optimierung der Anweisungen verwendet wurden. Dabei sollten insbesondere kritische Fehler, Grenzfälle, Ausfälle und unzulässige Aktionen betrachtet werden. Nach Änderungen an Modell, Prompts, Skills oder Werkzeugen sollten entsprechende Regressionstests wiederholt werden.

Info: Der Wert von Prompts

Ein Prompt, der über viele Iterationen entwickelt, an realen Fällen getestet und zuverlässig auf eine komplexe Aufgabe abgestimmt wurde, kann einen eigenen wirtschaftlichen Wert besitzen. In ihm können Erfahrungen darüber stecken, welche Informationen benötigt werden, welche Fehler häufig auftreten und wie Ergebnisse strukturiert oder geprüft werden sollen. Solche Prompts sollten daher ähnlich wie andere Prozessbeschreibungen, Vorlagen oder Konfigurationen dokumentiert, versioniert und gegebenenfalls vor unbefugtem Zugriff geschützt werden. Auch Anbieter von KI-Systemen verwenden interne Systemanweisungen zur Steuerung ihrer Modelle; deren genaue Ausgestaltung ist nicht immer öffentlich.

Ein guter Prompt sollte zunächst eindeutig benennen, welches Problem bearbeitet und welches Ergebnis erzeugt werden soll. Hilfreich ist außerdem relevanter Kontext, beispielsweise der Zweck einer Aufgabe, die Zielgruppe oder Informationen zur Ausgangssituation. Je genauer beschrieben wird, was benötigt wird und welche Grenzen gelten, desto weniger muss das Modell selbst über die gewünschte Vorgehensweise oder das erwartete Ergebnis schließen. Zu einem Prompt können deshalb auch Anforderungen an Umfang, Aufbau oder Ausgabeformat gehören. Ebenso kann festgelegt werden, welche Informationen berücksichtigt werden sollen und welche Handlungen nicht durchgeführt werden dürfen. Beispiele sind besonders dann hilfreich, wenn ein bestimmtes Format, eine bestimmte Schreibweise oder eine wiederkehrende Struktur erwartet wird. Es ist auch möglich dem Agenten eine Rolle zuzuweisen im Sinne von “Du bist ein Mitarbeiter im Einkauf”, allerdings lässt das durchaus viel Spielraum für Interpretationen. Dennoch kann es nützlich sein mit diesen die Ergebnisse zu verbessern und so evtl. einen Rückschluss zu ziehen welche Informationen für den eigentlichen Prompt noch förderlich sind.

Auch wenn ein guter Prompt maßgeblich für ein gutes Ergebnis ist, garantiert er dieses dennoch nie. Aussagen, Berechnungen und Schlussfolgerungen sollten weiterhin überprüfbar bleiben. Besonders wertvoll sind deshalb Aufgaben, bei denen der Agent seine verwendeten Quellen, Prüfschritte oder ausgeführten Aktionen nachvollziehbar dokumentieren kann. Gleichzeitig gilt für den bereitgestellten Kontext, dass nicht möglichst viele, sondern die für die Aufgabe notwendigen und zulässigen Informationen zur Verfügung gestellt werden sollten. Insbesondere personenbezogene oder vertrauliche Unternehmensdaten dürfen nur innerhalb einer dafür vorgesehenen technischen und organisatorischen Umgebung verarbeitet werden.

Bei agentischen Systemen verändert sich die Bedeutung des einzelnen Prompts zusätzlich. Wiederkehrende Regeln, Prüfkriterien und Verfahrensweisen müssen nicht vollständig in jedem Arbeitsauftrag enthalten sein. Sie können dauerhaft als Systemanweisungen, Skills, Referenzdokumente oder Werkzeuge, wie in diesem Kapitel gezeigt, bereitgestellt werden. Der Prompt beschreibt dann vor allem den aktuellen Auftrag, während die bereits konfigurierte Umgebung festlegt, wie der Agent diesen bearbeiten soll. Dadurch werden Arbeitsaufträge kürzer und die zugrunde liegenden Regeln lassen sich zentral pflegen.

4 SICHERHEITSASPEKTE IM UMGANG MIT AGENTISCHEN SYSTEMEN

Der Umgang mit agentischer KI kann grob in drei Kategorien gegliedert werden: Begrenzen von Handlungsmöglichkeiten, Angriffe von außen und Überwachung. Eine geeignete Orientierung für die Absicherung agentischer Systeme bietet das Open Web Application Security Project (OWASP) mit dessen *Top 10 for Agentic Applications 2026* [25]. Die Zusammenstellung entstand unter Beteiligung von mehr als 100 Fachleuten aus Forschung und Praxis und soll ausdrücklich als Ausgangspunkt für die Sicherheitsbewertung agentischer Anwendungen dienen. Für eine konkrete Implementierung lohnt sich deshalb der Blick in das vollständige Dokument, da OWASP zu jeder Risikoklasse typische Angriffsszenarien und Gegenmaßnahmen beschreibt.

Für eine konkrete Implementation sind nicht alle Risikoklassen gleichsam relevant. Ein einzelner Agent, der einen klar begrenzten Geschäftsprozess bearbeitet, wird beispielsweise keine Schnittpunkte mit komplexen Angriffen auf die Kommunikation zwischen vielen Agenten haben. Andere Klassen wie die Manipulation des Agentenziels sowie die Frage, welche Werkzeuge, Daten und Berechtigungen der Agent überhaupt besitzt, spielen für nahezu jeden Umgang mit agentischer KI eine wichtige Rolle.

Ein lokal betriebenes Sprachmodell bietet insbesondere bei vertraulichen Unternehmensdaten einen Sicherheitsvorteil, weil Eingaben und Ausgaben die eigene Infrastruktur für die Modellinferenz nicht verlassen müssen. Dieser Vorteil geht jedoch mit einer Verlagerung der Verantwortung einher: Absicherung der Infrastruktur, Zugriffskontrolle, Aktualisierung der Software und Prüfung der verwendeten Modelle liegen vollständig beim Unternehmen. Bei kommerziellen Diensten wie Claude oder Codex werden Daten dagegen durch eine externe Infrastruktur verarbeitet. Zwar stellen Enterprise-Angebote inzwischen umfangreiche Datenschutz-, Zugriffs- und Aufbewahrungskontrollen bereit, diese sind für einen Einsatz im Unternehmen jedoch als essenziell anzusehen; Dienste ohne entsprechende Schutzmechanismen sollten daher nicht in Betracht gezogen werden. Gleichzeitig übernimmt der Anbieter einen Teil der technischen Absicherung. Funktionen wie Sandboxing, Berechtigungssysteme und Freigaben gehören dabei nicht zum Sprachmodell selbst, sondern zum verwendeten Harness, sodass lokale und externe Modelle grundsätzlich mit unterschiedlichen Agentenumgebungen kombiniert werden können. Die Entscheidung zwischen lokalem und externem Modell ist daher weniger eine Entscheidung zwischen „sicher“ und „unsicher“ als zwischen unterschiedlichen Vertrauensgrenzen und Verantwortlichkeiten.

4.1 HANDLUNGSMÖGLICHKEITEN UND BERECHTIGUNGEN

Die Begrenzung der Handlungsmöglichkeiten ist für Unternehmen der wirksamste und unmittelbar kontrollierbare Sicherheitsmechanismus. Ein Agent kann Schaden verursachen, wenn er ein zulässiges Werkzeug falsch verwendet, einen Auftrag missversteht oder einen ungeeigneten Lösungsweg auswählt. Die Sicherheitsarchitektur sollte deshalb nicht davon ausgehen, dass der Agent jederzeit die richtige Entscheidung treffen wird. So sollten nur Werkzeuge bereitgestellt werden, die für die

jeweilige Aufgabe tatsächlich benötigt werden. Ein Agent zur Auswertung eines Funktionspostfachs benötigt beispielsweise zunächst Leserechte, aber keine Berechtigung zum Löschen oder Versenden von Nachrichten. Selbes gilt für eine Datenbankauswertung. Statt einen Agenten über die Rechte eines Mitarbeiters zu betreiben, ist es sinnvoller, ihm eine eigene technische Identität zu geben. Berechtigungen können auf diese Weise beschränkt und zeitlich begrenzt vergeben werden. Ein kompromittierter oder fehlgeleiteter Agent kann dadurch nur innerhalb eines deutlich kleineren Bereichs handeln, was den potenziellen Schaden minimiert. Das OWASP empfiehlt entsprechend Least Agency und Least Privilege: Rechte sollen nicht nur für den Agenten insgesamt, sondern möglichst für einzelne Werkzeuge, Datensätze und Aktionen begrenzt werden. Für besonders folgenreiche Aktionen können zusätzliche Authentifizierung, menschliche Freigaben oder eine vorherige Vorschau der geplanten Änderung verlangt werden.

Dass technische Grenzen auch ohne externen Angreifer notwendig sind, zeigte ein Vorfall mit dem Replit-Agent. Im Juli 2025 löschte der Agent bei einem Nutzer Daten aus einer Datenbank trotz ausdrücklich angeordnetem Code-Freeze. Möbus [26] berichtete, dass Schwierigkeiten und Bugs durch den Agenten verschwiegen und sogar falsche Daten und Reports kreiert wurden, um den Vorfall zu verschleiern. Erst im Anschluss wurde der Replitagent weiterentwickelt, um die Entwicklungsumgebung von der Produktion zu trennen. Das Unternehmen des Nutzers selbst verlor zum Glück nur ein wenig Arbeitszeit, da die Daten über die vorhandene Rollback-Funktion vollständig wiederhergestellt werden konnten.

4.2 EXTERNE ANGRIFFE

Die Arbeit mit Agenten ist, wie jede Arbeit am Computer mit Netzwerkzugang, gefährdet durch externe Intervention. Jedoch entstehen hier auch ganz neue Angriffsvektoren. Dadurch, dass Angriffe nun auch in natürlicher Sprache möglich sind, senkt dies teilweise die notwendigen Kenntnisse des Angreifers. Auch muss ein Angreifer gar keinen direkten Zugriff auf das System oder den Agenten besitzen. Anweisungen für einen Agenten können in Mails, einem Dokument oder auf einer Webseite versteckt sein. Das Modell erhält dann sowohl den eigentlichen Arbeitsauftrag als auch den manipulierten Inhalt und kann die dort enthaltenen Anweisungen fälschlicherweise als Teil seiner Aufgabe behandeln [25], [27]. Das Ziel des Angriffs ist oft ein Datenabfluss und die Manipulation angebundener Funktionen. Eine klare Trennung zwischen Daten und Anweisungen existiert für einen Agenten nicht. Modelle werden zwar darauf trainiert, Anweisungen des Systems und des Nutzers höher zu gewichten als Inhalte aus Dokumenten oder Werkzeugergebnissen; eine Garantie,dass es nicht dennoch zur Ausführung von externen Anweisungen kommt, stellt dies aber nicht dar.

Zur Abwehr von externen Angriffen wäre unter anderem ein Prozess, der verdächtige Inhalte überprüft, bevor es an den Agenten mit Anweisung zum Bearbeiten gegeben wird, denkbar. Eine Art Frühwarnsystem, ebenfalls umgesetzt durch einen Agenten, allerdings mit der konkreten Aufgabe, auf verdächtige Inhalte zu prüfen, statt diese zu nutzen. Ein solches Verfahren stellt jedoch keine verlässliche Sicherheitsgrenze dar, da auch der prüfende Agent selbst manipuliert werden kann. Die von OWASP, Bundesamt für Sicherheit in der Informationstechnik und dem National Cyber Security

Centre empfohlene Strategie ist jedoch kongruent mit dem ersten Abschnitt dieses Kapitels. Ein System, das schädlichen Anweisungen nicht Folge leisten kann, weil entsprechende Zugriffe oder Berechtigungen fehlen, stellt eine enorme Hürde für Angreifer dar [27], [28]. Das Problem hierbei liegt am Ende in der Flexibilität des Systems. Je allgemeiner ein Agent eingesetzt werden soll, desto mehr Datenquellen, Werkzeuge und Berechtigungen benötigt er typischerweise. Je enger diese Möglichkeiten begrenzt werden, desto kleiner wird dagegen die mögliche Schadenswirkung. Gleichzeitig steigt aber der Aufwand, wenn das System später für weitere Aufgaben eingesetzt werden soll.

4.3 ÜBERWACHUNG

Da eine fehlerfreie KI in absehbarer Zukunft unrealistisch ist, gibt es weitere Maßnahmen, die ergriffen werden können, um Risiken weiter zu minimieren. Die abschließende Strategie eines Systems lautet: Überwachung. Ob ein Agent während einer gängigen Aufgabe ungewöhnliche Aktivitäten durchführt, auf kritische Systeme zugreifen will, oder Ergebnisse offensichtliche Fehler enthalten, in vielen Fällen lässt sich eine Form der Überwachung einführen, die Auswirkungen begrenzt. Besonders problematisch sind lange Verkettungen von Aufgaben. Geschieht früh in der Kette ein Fehler, kann sich dieser nicht nur durch die Kette ziehen, sondern noch weiter verstärken. Um Fehler nachzuvollziehen, ist es hilfreich, den Agenten anzuweisen, Zugriffe und Schritte zu protokollieren. Das ist auch für die allgemeine Verbesserung des agentischen Systems eine gute Hilfestellung.
Bei der menschlichen Beteiligung am operativen Ablauf eines agentischen Systems lassen sich drei Formen unterscheiden: Human-out-of-the-Loop ohne unmittelbare menschliche Beteiligung, Human-on-the-Loop mit menschlicher Aufsicht und Eingriffsmöglichkeit sowie Human-in-the-Loop mit verpflichtender Prüfung oder Freigabe bestimmter Schritte. Technische Überwachung kann und sollte risikogerecht alle drei Formen begleiten. Ein Human-out-of-the-Loop-Betrieb kann dort eingeführt werden, wo Ergebnisse gut automatisch zu prüfen sind oder Aufgaben mit definierten Zugriffen verknüpft sind, die das Verhindern von Zugriffen an anderer Stelle einfach gestalten. In den meisten Fällen können diese Maßnahmen aber schon während der Implementationsphase als Werkzeuge oder über Rechte definiert werden. Bei dem Human-on-the-Loop-Ansatz kann ein Mensch in einen grundsätzlich bereits automatisierten Prozess eingreifen.
Gerade aber weil es in einem solchen Prozess an keinem Punkt zwingend notwendig ist, kann diese Aufgabe schnell monoton werden, was sich auf die Aufmerksamkeit und die Überwachung selbst auswirken kann [29]. Beim *Human-in-the-Loop*-Ansatz muss ein Mensch bestimmte Ergebnisse oder geplante Aktionen prüfen und freigeben, bevor der Agent fortfahren darf. Dieser eignet sich besonders für Schritte mit hohen oder nur schwer rückgängig zu machenden Folgen, etwa das Versenden verbindlicher Mitteilungen, das Löschen von Daten oder Änderungen an produktiven Systemen. Der zusätzliche Kontrollschritt reduziert zwar den Automatisierungsgrad und verursacht weiterhin Arbeitsaufwand, kann aber verhindern, dass einzelne Fehlentscheidungen unmittelbar wirksam werden. Die Freigabe sollte sich deshalb auf wenige tatsächlich kritische Punkte konzentrieren und dem Mitarbeiter ausreichend Informationen zur Beurteilung der geplanten Aktion bereitstellen. Es ist

auch möglich, solche Freigaben in einen Testlauf auf einer unkritischen Instanz zu schicken, falls die Auswirkungen der Freigabe nicht gut vorherzusehen sind. In diesem Ansatz arbeiten Agent und Mitarbeiter am ehesten Hand in Hand. Der Mitarbeiter entscheidet, welche Aufgaben der Agent übernehmen soll und welche er lieber selbst übernimmt. Dadurch wird der Großteil des Risikos gar nicht erst auf das agentische System gelegt. Bei sehr vielen Freigaben droht auch hier Ermüdung und unkritisches Durchwinken (Automation Bias).

Eine gute Praxis im Umgang mit Überwachung sollte für mindestens alle zentralen Aufgaben eines Prozesses definieren, in welche der Überwachungsstufen diese fallen. Innerhalb eines Prozesses können dabei unterschiedliche Stufen sinnvoll sein. Zusätzlich sollte festgelegt werden, unter welchen Bedingungen ein Vorgang eskaliert wird, wer für die anschließende Prüfung verantwortlich ist und innerhalb welcher Zeit ein Eingriff erfolgen muss. Warnungen, Freigaben und Eingriffe sollten nachvollziehbar protokolliert werden. Ändern sich der Prozess, die eingesetzten Werkzeuge oder die Rechte des Agenten, sollte auch die gewählte Form der Überwachung erneut geprüft werden.

5 ZUSAMMENFASSUNG

Die Einführung agentischer Systeme in Unternehmensprozesse ist bereits in vollem Gang, auch wenn die Technologie noch jung und viele Anwendungen weiterhin experimentell sind. Theoretisch lassen sich mit agentischer KI große Teile betrieblicher Prozesse automatisieren. In der Praxis kann die Einführung jedoch technisch aufwendig, risikoreich und rechtlich komplex sein. Wirtschaftlich interessante Prozesse können dennoch lohnenswert sein. Vor allem dort, wo hauptsächlich Informationen aus unterschiedlichen Quellen verarbeitet und standardisierte Dokumente erstellt werden müssen. Eine Betrachtung des zu erwartenden Integrationsaufwands, der rechtlichen Anforderungen sowie möglicher Folgen von Fehlern wird vorab dringend empfohlen.

Um einen Prozess mittels agentischer KI umzusetzen, müssen Aufgaben in geeignete Arbeitsschritte zerlegt und teilweise vom Mitarbeiter auf das System übertragen werden. Dazu benötigt der Agent neben einem konkreten Arbeitsauftrag geeignete Datenquellen, Werkzeuge, Berechtigungen, Regeln und gegebenenfalls wiederverwendbare Skills. Diese Bestandteile können am Ende selbst einen Unternehmenswert darstellen. Die Entwicklung erfolgt in der Regel iterativ. Erste Ergebnisse sind im Detail oft nicht zufriedenstellend, auch wenn die äußere Form dies zunächst schnell vermuten lässt. Mitarbeiter, die den Prozess gut kennen, justieren dann die Bestandteile der jeweiligen Aufgaben, bis auch die Details zufriedenstellende Ergebnisse zeigen. Mitarbeiter verschwinden dabei nicht zwangsläufig am Ende des Prozesses. Ihre Aufgaben verschieben sich jedoch von der Durchführung einzelner Routinetätigkeiten stärker hin zur Prüfung, Freigabe und Pflege der Wissensbasis und Behandlung ungewöhnlicher Fälle.

Mit der Einführung von agentischen Systemen entstehen auch neue Angriffsvektoren und Sicherheitsrisiken in Unternehmensprozessen. Die effektivste Methode, um diesen zu begegnen, sind die Einschränkungen von Berechtigungen und Handlungsmöglichkeiten. Nicht nur verhindert dies den Zugriff auf Systemteile, die für die jeweiligen Prozesse nicht benötigt werden, sondern verringert auch die möglichen Auswirkungen eines externen Angriffs und die Notwendigkeit von Überwachung. Im Gegenzug wird dadurch die Flexibilität im Einsatz des Agenten eingeschränkt, was durch den Einsatz unterschiedlicher Agenten ebenfalls abgemindert werden kann.

Für Beratung und Hilfe bei der Umsetzung gibt es bundesweit einige Anlaufstellen. Unter anderem die Mittelstand-Digital-Zentren und die Zukunftszentren mit teilweise kostenfreien und anbieterneutralen Beratungsangeboten. In Nordrhein-Westfalen bietet zusätzlich KI.NRW Formate von der ersten Beratung über die Identifikation geeigneter Anwendungsfälle bis zur Ausarbeitung konkreter Umsetzungskonzepte an. European Digital Innovation Hubs ermöglichen darüber hinaus den Zugang zu technischem Fachwissen und die Erprobung digitaler Technologien vor größeren Investitionen. Für die anschließende technische Umsetzung können spezialisierte Forschungs-, Beratungs- oder Integrationspartner hinzugezogen werden.

Literaturverzeichnis

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, Curran Associates, Inc., 2012. Available: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[3] A. Vaswani *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, Curran Associates, Inc., 2017. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[4] T. B. Brown *et al.*, "Language models are few-shot learners," in *Advances in neural information processing systems*, Curran Associates, Inc., 2020, pp. 1877–1901. Available: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[5] R. Bommasani *et al.*, "On the opportunities and risks of foundation models." 2021. Available: https://arxiv.org/abs/2108.07258

[6] M. Abou Ali, F. Dornaika, and J. Charafeddine, "Agentic AI: A comprehensive survey of architectures, applications, and future directions," *Artificial Intelligence Review*, vol. 59, no. 1, p. 11, 2025.

[7] A. Bandi, B. Kongari, R. Naguru, S. Pasnoor, and S. V. Vilipala, "The rise of agentic ai: A review of definitions, frameworks, architectures, applications, evaluation metrics, and challenges," *Future Internet*, vol. 17, no. 9, p. 404, 2025.

[8] S. Yao *et al.*, "ReAct: Synergizing reasoning and acting in language models," in *Proceedings of the 11th international conference on learning representations*, 2023. Available: https://openreview.net/forum?id=WE_vluYUL-X

[9] S. Russell and P. Norvig, *Artificial intelligence: A modern approach*, 4th ed. Hoboken, NJ: Pearson, 2021.

[10] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, 2000, doi: 10.1109/3468.844354.

[11] K. J. K. Feng, D. W. McDonald, and A. X. Zhang, "Levels of autonomy for AI agents." 2025. Available: https://arxiv.org/abs/2506.12469

[12] D. Safin and D. Balta, "Autonomy and agency in agentic AI: Architectural tactics for regulated contexts." 2026. Available: https://arxiv.org/abs/2605.12105

[13] SAE International, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," SAE International, Recommended Practice J3016_202609, Sep. 2026. doi: 10.4271/J3016_202609.

[14] Microsoft, "Supplier communications features of the procurement agent (production-ready preview)." Microsoft Learn, Dynamics 365 Supply Chain Management, 2026. Accessed: Oct. 03, 2026. [Online]. Available: https://learn.microsoft.com/en-us/dynamics365/supply-chain/procurement/procurement-agent-supplier-com-overview

[15] OpenClaw, "Heartbeat." OpenClaw Documentation, 2026. Available: https://docs.openclaw.ai/gateway/heartbeat

[16] Letta, “Sleep-time agents.” Online-Dokumentation, 2026. Accessed: Oct. 03, 2026. [Online]. Available: https://docs.letta.com/guides/agents/architectures/sleeptime

[17] M. Zeff, “OpenAI launches ChatGPT pulse to proactively write you morning briefs.” TechCrunch, 2025. Accessed: Oct. 03, 2026. [Online]. Available: https://techcrunch.com/2025/09/25/openai-launches-chatgpt-pulse-to-proactively-write-you-morning-briefs

[18] Nous Research, “Skills system.” Online-Dokumentation, 2026. Accessed: Oct. 03, 2026. [Online]. Available: https://hermes-agent.nousresearch.com/docs/user-guide/features/skills

[19] OpenClaw, “Skill workshop.” Online-Dokumentation, 2026. Accessed: Oct. 03, 2026. [Online]. Available: https://docs.openclaw.ai/tools/skill-workshop

[20] Y. Nakajima, “BabyAGI (archive).” GitHub-Repository, 2024. Accessed: Oct. 03, 2026. [Online]. Available: https://github.com/yoheinakajima/babyagi_archive

[21] R. Lakshmanan, “Researchers find 341 malicious ClawHub skills stealing data from OpenClaw users.” The Hacker News, 2026. Accessed: Oct. 03, 2026. [Online]. Available: https://thehackernews.com/2026/02/researchers-find-341-malicious-clawhub.html

[22] Intercom, “Manage fin AI agent’s escalation guidance and rules.” Intercom Help Center, 2026. Accessed: Oct. 03, 2026. [Online]. Available: https://www.intercom.com/help/en/articles/12396892-manage-fin-ai-agent-s-escalation-guidance-and-rules

[23] Europäische Union, “Konsolidierte Fassung der Verordnung (EU) 2024/1689 über künstliche Intelligenz.” EUR-Lex, Fassung vom 27. Juli 2026, 2026. Accessed: Sep. 17, 2026. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/2026-07-27

[24] J. Bager, “Kontrollierte Intelligenz: Wo KI helfen kann und in welchen Fällen andere Werkzeuge geeigneter sind,” *c’t*, vol. 14, pp. 60–65, 2026.

[25] OWASP Gen AI Security Project, “OWASP top 10 for agentic applications 2026,” OWASP Gen AI Security Project – Agentic Security Initiative, Report, Dec. 2025. Accessed: Sep. 02, 2026. [Online]. Available: https://genai.owasp.org/download/52117/?tmstv=1765059207

[26] M. Möbus, “Künstliche intelligenz: Vibe-coding-dienst replit löscht produktionsdatenbank.” heise online, Jul. 25, 2025. Accessed: Sep. 02, 2026. [Online]. Available: https://www.heise.de/news/Kuenstliche-Intelligenz-Vibe-Coding-Dienst-Replit-loescht-Produktionsdatenbank-10499478.html

[27] Bundesamt für Sicherheit in der Informationstechnik, “Generative KI-modelle: Chancen und Risiken für Industrie und Behörden,” Bundesamt für Sicherheit in der Informationstechnik, Bonn, Jan. 2025. Accessed: Sep. 02, 2026. [Online]. Available: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Generative_KI-Modelle.pdf?__blob=publicationFile&v=5

[28] National Cyber Security Centre, “Managing the cyber risk of agentic AI.” NCSC Blog, Aug. 2026. Accessed: Sep. 11, 2026. [Online]. Available: https://www.ncsc.gov.uk/blogs/managing-the-cyber-risk-of-agentic-ai

[29] J. Gouraud, A. Delorme, and B. Berberian, “Autopilot, mind wandering, and the out of the loop performance problem,” *Frontiers in neuroscience*, vol. 11, p. 541, 2017.

IMPRESSUM

HERAUSGEBER

INHALTLICHE VERANTWORTUNG

Frochte, Jörg[1]

Priebe, Sebastian[1]

[1]Hochschule Bochum, Interdisziplinäres Institut für Angewandte KI und Data Science Ruhr (AKIS)

GESCHÄFTSSTELLE TRAIBER.NRW

Bergische Universität Wuppertal
Institute for Technologies and Management for Digital Transformation (TMDT)

Gebäude FZ | Ebene 01 | Raum 19
Lise-Meitner-Str. 27-31, 42119 Wuppertal
Telefon: 0202 439-1164
E-Mail: koordination@traiber.nrw
www.traiber.nrw

Wuppertal, Oktober 2026

Gefördert durch:

Bundesministerium für Wirtschaft und Energie

aufgrund eines Beschlusses des Deutschen Bundestages

TRAIBER.NRW

TRANSFORMATION
DER AUTOMOTIVE INDUSTRIE
IN DER BERGISCHEN REGION

Gefördert durch:

Bundesministerium
für Wirtschaft
und Energie

aufgrund eines Beschlusses
des Deutschen Bundestages