

Der CIO-Leitfaden für agentenbasiertes Programmieren: Unternehmensanwendungen mit KI-Geschwindigkeit und absoluter Zuverlässigkeit entwickeln

Ein Leitfaden für COOs, CXOs
und das operative Management



Der Druck ist real – das Problem auch

Die Frage, ob künstliche Intelligenz implementiert werden sollte, ist längst kein Vorstandsthema mehr. Mittlerweile interessiert den Vorstand nur noch, wann und wie schnell KI verfügbar ist sowie ob sich die Technologie rechnet. Für COOs und operative Führungskräfte führt dieser Druck zu zahlreichen Pilotprojekten mit autonomen Agenten – also Implementierungen, bei denen KI mit minimaler menschlicher Beteiligung Kundeninteraktionen übernimmt, Vorgänge weiterleitet und Workflows koordiniert.

Die ersten Ergebnisse waren vielversprechend: schnellere Lösungen, geringerer Personalbedarf und weniger manuelle Eskalationen. Doch beim Skalieren dieser Pilotprojekte in die Produktion ergibt sich ein anderes Bild: Die Kosten entwickeln sich nicht wie prognostiziert, die Ergebnisse sind nicht so konsistent, wie die Demos nahelegen, und die operative Kontrolle, von der Führungskräfte ausgegangen sind, wird zusehends untergraben.

56 % der CEOs verzeichnen derzeit keinen messbaren finanziellen Vorteil durch KI-Investitionen ([PwC 29th Global CEO Survey, Januar 2026](#)). Nicht die Technologie ist hier das Problem, sondern die Architektur. Die Frage ist nicht, ob KI die operative Effizienz verbessert, sondern ob heutige KI-Implementierungen in vielen Unternehmen die richtige Struktur mitbringen, damit KI ihr volles Potenzial entfalten kann.

Die Antwort lautet meistens „Nein“.



Die versteckten Reasoning-Kosten

KI hat die Wirtschaftlichkeit der Unternehmensautomatisierung grundlegend verändert. Token – die Maßeinheit für die Rechenleistung und das „Denken“ eines KI-Systems – haben sich zu einem der wichtigsten Budget-Posten bei der digitalen Transformation von Unternehmen entwickelt. Und in den meisten Unternehmen entwickelt sich dieser Posten auf eine Weise, mit der niemand gerechnet hat.

Als die Unternehmensleitung die Implementierung von KI-Agenten genehmigte, war das Kostenmodell noch simpel: Manuelle Arbeit wird durch automatisiertes Reasoning ersetzt, die Personalkosten werden gesenkt und die Marge kann sich sehen lassen. Dabei wurde jedoch nicht berücksichtigt, wie unvorhersehbar der Token-Verbrauch beim unternehmensweiten KI-Einsatz wird. Diese Unvorhersehbarkeit wird durch zwei Faktoren bestimmt, die zusammen den erwarteten ROI schmelzen lassen.

Der erste Faktor ist die Eigendynamik von KI-Agenten, deren Token-Verbrauch sich sukzessive akkumuliert. Das liegt daran, dass KI-Agenten im Dauereinsatz nicht nur einmalig ein Reasoning durchführen und dann zum nächsten Schritt übergehen, sondern für jeden Prozess-Schritt das Reasoning wiederholen und erneut den Kontext überprüfen, Optionen bewerten und Ergebnisse liefern. Der Token-Verbrauch für den Betrieb dieser Systeme summiert sich – zusätzlich zu den Token, die die Entwicklung von KI-Agenten kostet. Bei einem Einsatz von KI-Agenten im mittleren Umfang steigen deshalb sich die Kosten schneller, als es die meisten Finanzierungsmodelle vorhersagen.

Wir bezeichnen das als die Token-Steuer.

Der zweite Faktor ist die Preisinstabilität. Anfangs gab es bei KI-Anbietern subventionierte Flatrate-Preismodelle, die Unternehmen eine planbare, verlässliche Kostenstruktur für die Budgetierung garantierten. Diese Zeiten sind bald vorbei. Da die Ressourcen für die KI-Wirtschaft durch Engpässe bei Chips, Energie und Rechenzentrumskapazitäten knapp werden und KI-Anbieter vor dem Börsengang stehen, müssen Unternehmen ständige Preisanpassungen hinnehmen, bei denen bisherige Flatrates zunehmend einer verbrauchsabhängigen Abrechnung weichen. Das bedeutet, dass die Kosten für ein und denselben Workload stark variieren können. Sam Altman räumte im Juni 2026 ein, dass die KI-Token-Kosten selbst für KI-Anbieter zu einem „großen Problem“ geworden sind ([Axios, Juni 2026](#)).

Die Folge ist, dass Unternehmen ihr Budget für die digitale Transformation trotz ungewissem ROI förmlich verbrennen. Gartner geht davon aus, dass über 40 % der agentenbasierten KI-Projekte bis Ende 2027 abgebrochen und eskalierende, unvorhersehbare Betriebskosten als Hauptgrund genannt werden ([Gartner, Juni 2025](#)).

Die Chefetage stellt heute nicht den Einsatz von KI in Frage, möchte jedoch wissen, wie sich ein positiver ROI erzielen lässt, wenn sich die zugrunde liegende Kostenstruktur ständig ändert. Dies lässt sich nur mit einem Perspektivwechsel lösen: Statt für Tokens zu zahlen, sollte in Ergebnisse investiert werden. Konkret bedeutet das den Aufbau einer Architektur, bei der die Ausgaben für KI-Token den Aufgaben vorbehalten sind, die tatsächlich ein Reasoning erfordern. Alles andere – also Aufgaben ohne Reasoning – werden dagegen mit einer kalkulierbaren, fixen Kostenstruktur finanziert. So lässt sich vermeiden, dass die Kosten regelmäßig das Budget sprengen.

Wenn KI „kreativ“ wird

Unvorhersehbare Kosten sind nur ein Aspekt des Problems. Der andere Aspekt ist: Was passiert, wenn KI-Agenten in Bereichen autonom entscheiden dürfen, in denen eine Automatisierung fatal sein kann?

Große Sprachmodelle (Large Language Models, LLMs) sind hilfreich und erstellen Antworten, die gut klingen – dafür wurden sie schließlich trainiert. Wofür sie aber nicht trainiert wurden, ist die Befolgung Ihrer Compliance-Richtlinien. LLMs verstehen nicht automatisch Ihre SLAs, Ihre regulatorischen Verpflichtungen, Ihre Eskalationsregeln oder die Kundenversprechen Ihres Unternehmens. Wenn autonome Agenten ohne strenge Grenzen Workflows mit direktem Kundenkontakt übernehmen, kann sich das, was die KI für eine sinnvolle Antwort hält, und das, was Ihr Compliance-Team genehmigen würde, erheblich unterscheiden.

Das ist keine Frage der Risikotoleranz, sondern der Architektur.

Die Entscheidungen mit dem größten rechtlichen und Reputationsrisiko – wie Klärung von Anspruchsberechtigungen, Streitbeilegungen, RichtlinienAusnahmen oder regulierte Kommunikation – sind zugleich auch jene, für die es eine klar definierte richtige Antwort gibt. Diese Entscheidungen sind nicht mehrdeutig und profitieren auch nicht von probabilistischen Schlussfolgerungen. Die Weiterleitung eines Vorgangs basierend auf dem Kontostatus, die Prüfung, ob ein Schadensfall eine dokumentierte Checkliste erfüllt, oder die Anwendung einer RichtlinienAusnahme nach definierten Kriterien – all dies sind deterministische Abläufe. Überlässt man das alles dem Reasoning eines KI-Agenten, bietet das keinen Mehrwert. Stattdessen werden Prozesse, die Konsistenz erfordern, und Interaktionen, bei denen es keine Fehler geben darf, plötzlich variabel.

KI-Halluzinationen sind in großen Produktionsumgebungen keine Seltenheit. Warum? Weil genau das passieren muss, wenn Systeme Antworten liefern, die auf probabilistischer Inferenz statt auf deterministischen Regeln basieren. In einem unwichtigen Kontext ist eine Halluzination vielleicht ärgerlich. Bei Kundenbeschwerden, Workflows für Anspruchsberechtigungen, bei Interaktionen im Finanzwesen oder der Kommunikation in einer regulierten Branche entsteht daraus ein Haftungsrisiko.

Zu den wichtigsten Gartner-Vorhersagen für 2026 gehört die „Death by AI“-Prognose. Demnach wird mit weltweit mehr als 2.000 Klagen wegen Vorfällen gerechnet, die aufgrund mangelnder Governance und Risikokontrollen bei KI-generierten Ergebnissen entstehen ([Gartner, Oktober 2025](#)). Das sind keine abstrakten Risiken für zukünftige Implementierungen, es ist eine akute Gefahr für jedes Unternehmen, das autonomen KI-Agenten eine weitreichende Befugnis für regulierte oder Workflows mit direktem Kundenkontakt erteilt hat.

Die produktivere Lösung besteht in einer Architektur mit Geschäftsregeln dafür, was KI-Agenten dürfen – und was nicht. KI sollte niemals den Datenbestand Ihrer Compliance-Logik ersetzen, sondern muss stets innerhalb der vorgegebenen Compliance-Grenzen arbeiten. Die richtige Vorgehensweise bei glasklaren Entscheidungen wird durch die Business-Logik bestimmt und auf der Orchestrierungs-Ebene durchgesetzt – sie wird nicht an ein Modell delegiert, dem am nächsten Tag vielleicht eine ganz andere logische Antwort einfällt. Gibt es eine solche Trennung, wird die Compliance strukturell und ist nicht nur eine Aufsichtsfunktion. Anders ausgedrückt: Die KI kann nicht „kreativ“ werden und von Ihren Vorgaben abweichen, weil Ihre Vorgaben nicht zur Diskussion stehen.

Die Architektur, die beides bietet

Die operative Herausforderung, vor der die meisten Unternehmen stehen, wird nicht durch zu viel KI verursacht, sondern durch den Einsatz von KI in den falschen Prozessarchitektur-Bereichen.

Die wirtschaftlichen Gründe für den KI-Einsatz in Unternehmen liegen auf der Hand: Senkung der Servicekosten, Klärung von mehr Kundenanliegen und einheitlichere Kundenerlebnisse unternehmensweit. Dies sind erreichbare Ergebnisse, aber sie verlangen einen unverstellten Blick darauf, wo KI von Wert ist und wo dadurch ohne merkbliche Rendite nur Kosten und Risiken entstehen.

Bei einer Architektur, die diese Ergebnisse durchweg liefert, arbeiten zwei Ebenen zusammen. Die erste ist die deterministische Orchestrierungs-Ebene: eine regelbasierte Engine, die Prozessabläufe koordiniert,

Business-Logik durchsetzt, den Status verfolgt und Arbeit nach vorgegebenen Kriterien weiterleitet. Diese Ebene verursacht nur marginale Kosten pro Transaktion. Hier findet kein Reasoning statt, sondern nur die Ausführung. Was auf dieser Ebene geschieht, ist komplett planbar, vollständig überprüfbar und ohne Halluzinationen.

Die zweite Ebene ist die KI-Inferenz, die präzise an den Prozesspunkten aufgerufen wird, an denen deterministische Logik nicht ausreicht: Analyse der Kundenstimmung, Interpretation von unstrukturierten Dokumenten, Antworten in natürlicher Sprache oder die Zusammenfassung von komplexen Vorgängen. Bei diesen probabilistischen Aufgaben wird KI zu einem Feature (und nicht zu einem Haftungsrisiko), weil es für die Art von Aufgaben keine einzige richtige Antwort gibt, die sich mit Regeln kontrollieren lassen würde.

Werden diese beiden Ebenen richtig voneinander getrennt, sinkt die Token-Steuer erheblich. Token werden dann hauptsächlich für die Aufgaben verbraucht, die Inferenz erfordern. Alles andere wird deterministisch ausgeführt. Dies senkt zugleich das Compliance-Risiko, weil die Orchestrierungs-Ebene die Geschäftsregeln unabhängig von den Ergebnissen der KI-Ebene durchsetzt. Die KI arbeitet damit innerhalb vorgegebener Grenzen, die eingehalten werden müssen.

Nur 12 % der CEOs geben an, dass KI gleichzeitig messbare Kostensenkungen und Umsatzsteigerungen bringt ([PwC, Januar 2026](#)). Diese 12 % implementieren nicht weniger KI als vergleichbare Unternehmen. Was sie jedoch unterscheidet, ist die Implementierung mit struktureller Disziplin.



Kontrolle ist der Wettbewerbsvorteil

Die operativen Führungskräfte, die mit KI dauerhaft einen ROI erzielen werden, sind nicht diejenigen, die die Technologie am schnellsten implementieren oder bei Agenten auf maximale Autonomie setzen. Ihr Erfolgsrezept besteht darin, die Kontrolle über die Prozessarchitektur zu behalten und KI genau dort zu integrieren, wo sie wirklich einen Wert bringt.

Die Orchestrierung wird damit zu einem strategischen Element, nicht zu einer Legacy-Beschränkung. Das bedeutet, dass klar definiert wird, welche Entscheidungen deterministisch sind und welche Inferenz erfordern. Auf diesen Vorgaben wird dann eine Architektur aufgebaut, die diese Grenzen konsequent durchsetzt. Das bedeutet auch, dass die KI-Kosten auf Transaktionsebene und nicht nur auf Plattformebene gemessen werden. So wird die Token-Steuer sichtbar, bevor sie das Budget sprengt.

Vorstände liegen richtig in der Annahme, dass KI die Servicekosten senken und das Kundenerlebnis verbessern kann. Doch dieses Ziel erreichen Unternehmen nicht mit autonomen Agenten, die den richtigen Workflow selbst herausfinden müssen. Die Lösung ist eine geregelte, orchestrierte Prozessarchitektur, bei der KI wie ein Präzisionsinstrument eingesetzt wird.

Sind Sie bereit, die Kontrolle zu übernehmen? Finden Sie mit dem [TCO-Rechner](#) heraus, wie viel Sie bei den Token-Kosten sparen können, wenn Sie für Workflows mit dem höchsten Arbeitsaufkommen auf einen strukturierten Orchestrierungsansatz setzen.





Vielen Dank!

Über uns

Pega bietet die führende KI-gestützte Plattform für die Transformation von Unternehmen. Viele der weltweit einflussreichsten Unternehmen vertrauen auf unsere Technologie zur Neugestaltung von Arbeitsabläufen – durch Workflow-Automatisierung, persönliche Kundenerlebnisse und modernisierte Legacy-Systeme. Seit 1983 treibt die skalierbare, flexible Architektur kontinuierlich Innovationen voran und hilft Kunden dabei, ihren Weg zum autonomen Unternehmen zu beschleunigen. **[PEGA.COM/DE](https://www.pega.com/de)**