



Versteckte KI-Kosten liegen nicht im Modell, sondern in der Umsetzung

Die Vorteile vorhersehbarer KI

Juni 2026 | Don Schuerman, CTO & Head of Marketing, Pega



Bei der nächsten KI-Phase geht es nicht um Möglichkeiten, sondern um Verantwortlichkeit.

In den letzten zwei Jahren hat sich die Unternehmens-KI von einer experimentellen zu einer praxistauglichen Technologie entwickelt: Projekte, die früher Expertenteams und lange Entwicklungszeiten erforderten, lassen sich heute in wenigen Tagen erledigen und schneller als je zuvor können Teams Code generieren, einsatzbereite Agenten entwickeln und Workflows automatisieren.

Diese Fortschritte haben auch die Perspektive der Unternehmensleitung verändert.

Noch vor einem Jahr fragten Führungskräfte: *Was kann KI für uns tun?*

Heute ist die Frage direkter: *Was kostet uns KI und welchen Wert bringt uns das?*

Je mehr Unternehmen von Pilotprojekten zum skalierten KI-Einsatz im operativen Geschäft übergehen, desto höher werden die Erwartungen. KI ist nicht länger an Innovationsbudgets gebunden, sondern wird zum Teil dessen, wie das Unternehmen funktioniert. Und das bedeutet, dass KI der kritischen Überprüfung ihrer Zuverlässigkeit, Governance und Kosten standhalten muss.

Alle Branchen begreifen zusehends die Chancen mit KI, anders sieht es aber mit der Wirtschaftlichkeit aus. Wie eine aktuelle Gartner-Studie¹ zeigt, sind sinnvolle Ergebnisse bei den meisten KI-Initiativen weiterhin die Ausnahme – während viele Unternehmen kaum bis gar keinen messbaren ROI aus früheren Investitionen sehen.

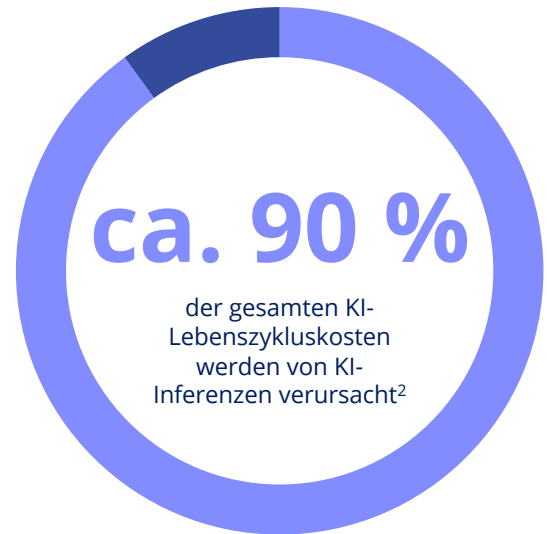
Das liegt nicht daran, dass KI nichts bringt. Es ist eher ein Signal, dass wir eine neue Phase betreten, in der Erfolg weniger davon abhängt, was Sie entwickeln können, sondern davon, wie effektiv Sie damit unternehmensweit arbeiten.

¹ Gartner-Studie zum Wert von KI (gartner.com/en/articles/ai-value)

Die KI-Entwicklung wird einfacher, ihr effizienter Einsatz schwerer

Bei der ersten KI-Welle ging es vor allem darum, die Technologie überhaupt zu nutzen. Mit der breiten Verfügbarkeit von leistungsstarken Modellen stürzten sich Unternehmen in Experimente, was Innovationen massiv beschleunigte. Beim Einsatz dieser Innovationen in der Praxis folgt jetzt die bittere Realität.

Die Kosten, die in der Pilotphase kalkulierbar schienen, steigen zunehmend, die Qualität der Ergebnisse schwankt und in vielen Fällen ist der geschäftliche Nutzen durch KI-Investitionen schwer zu messen. Im Kern dieser Herausforderung steht ein strukturelles Problem: Die meisten heutigen KI-Systeme wurden für **Experimente, nicht für einen verlässlichen Einsatz in der Praxis entwickelt**.



Viele basieren stark auf einem *Reasoning zur Laufzeit*, bei dem Modelle für jede Interaktion bei Null anfangen. Diese Flexibilität kann zwar nützlich sein, hat jedoch ihren Preis: Jede Interaktion verbraucht Rechenleistung – die mit zunehmender Komplexität stetig steigt – und je stärker KI genutzt wird, desto höher werden die Kosten.

Jedes Mal, wenn die KI eine Aufgabe neu bewerten, Kontext interpretieren oder über den nächsten Schritt entscheiden muss, entstehen zusätzliche Kosten. Wird KI in kleinem Umfang eingesetzt, lässt sich das im Griff behalten, doch beim skalierten Unternehmenseinsatz steigen die Kosten, weil das System nicht nur Aufgaben erledigt, sondern die gleichen Denkprozesse immer und immer wieder von vorn beginnt. Das hat dazu geführt, dass viele Unternehmen dies mittlerweile als eine Art „Token-Steuer“ empfinden, bei der jeder weitere Reasoning-Schritt zwar geringe Kosten verursacht, die sich aber durch die Skalierung summieren.

Wird KI im Unternehmen skaliert, ändert sich die Kostengleichung. In vielen Unternehmensumgebungen übersteigen die Inferenzkosten – also die Kosten für den kontinuierlichen Betrieb von Modellen in der Produktion – schnell die ursprünglichen Entwicklungskosten, wodurch KI von einer einmaligen Investition zu einer laufenden Betriebsausgabe wird.

Nicht der Einsatz von KI verursacht die wahren Kosten, sondern die Art und Weise, wie sie verwendet wird. Unternehmen zahlen unablässig dafür, die KI Aufgaben definieren zu lassen, die vorab hätten strukturiert werden können. Aus diesem Grund stimmt die KI-Leistung bei kontrollierten Pilotprojekten, geht jedoch bei der effektiven Skalierung für den unternehmensweiten Einsatz in die Knie. **Das Problem ist nicht das Potenzial von KI, sondern ihr effizienter Einsatz.**

² MIT Technology Review, „AI's energy usage and climate footprint“ (2025)

KI-Kosten hängen davon ab, wie wir KI einsetzen

Wenn Führungskräfte KI-Kosten in den Griff bekommen wollen, werden oft zuerst die Modelle unter die Lupe genommen: Man vergleicht Anbieter, Funktionalitäten und Preisstrukturen.

Aber die wichtigere Frage lautet: **Wofür setzen Sie KI im Unternehmen eigentlich ein? Und wie ist diese Arbeit strukturiert?**

In vielen Unternehmen arbeiten heute verschiedenste Teams mit KI, die alle unterschiedliche Modelle, Agenten und Workflows nutzen – der KI-Einsatz wird dabei kaum koordiniert. Das führt nicht nur zu mehr Komplexität, sondern auch zu doppelter Arbeit – für Teams und für die KI. Die gleichen Entscheidungen werden mehrmals an verschiedenen Stellen und oft mit einer anderen Logik getroffen.

Laut aktuellen Prognosen wird die KI-Nutzung in Unternehmen in den nächsten Jahren zu einem mehr als 20-fachen Anstieg des Token-Verbrauchs führen.³ Werden Agenten unternehmensweit skaliert, steigt nicht nur der KI-Einsatz, sondern auch dessen Kosten. Folglich wird KI oft als Ergänzung zu bestehenden Prozessen hinzugefügt, ohne die Funktionsweise dieser Prozesse grundlegend zu überarbeiten.

Das bringt vielleicht einige Verbesserungen, hat aber mit einer Transformation wenig zu tun und führt zudem zu zwei systemischen Problemen:

1. Die Ineffizienz bleibt: Ist der zugrunde liegende Prozess fragmentiert oder zu komplex, beseitigt der Einsatz von KI diese Komplexität nicht, sondern beschleunigt lediglich den Prozess.
2. Alles entscheidet sich zur Laufzeit: Statt vorher festzulegen, wie eine Aufgabe erledigt werden soll, überlassen Unternehmen es der KI, dynamisch über den nächsten Schritt zu entscheiden – Interaktion für Interaktion.

Diese Flexibilität wird vielleicht als leistungsstarke Option empfunden, führt aber auch zu Variabilität, steigenden Kosten und schwerer kontrollierbaren Ergebnissen. Letztlich ist das ein Problem der Architektur:

- Gut konzipierte Systeme verringern den Aufwand, bevor die eigentliche Arbeit beginnt.
- Schlecht strukturierte Systeme verursachen dagegen bei jeder Ausführung neue Kosten.

Wird KI eingesetzt, ohne sich vorher anzusehen, wie die Arbeit über verschiedene Systeme und Teams hinweg abläuft, steigen die Kosten mit der zunehmenden KI-Nutzung und werden sich auch nicht stabilisieren, wenn KI-Einsatz skaliert wird.

Deshalb steigen in vielen Unternehmen die Kosten schneller, als dass die KI sinnvolle Ergebnisse bringt.

³ Goldman Sachs, „AI agents forecast to boost tech cash flow as usage soars“

Probleme mit Prozessen werden durch KI nicht gelöst, sondern nur skaliert

Häufig wird KI einfach nur zu bestehenden Prozessen hinzugefügt: Man nimmt einen Workflow – z. B. im Kundenservice, beim Onboarding oder zur Bearbeitung von Versicherungsansprüchen – und ergänzt ihn mit einer KI-Ebene, die das Ganze verbessern soll. Das Benutzererlebnis wird dialogorientierter und das System responsiver,

aber der zugrunde liegende Prozess bleibt wie gehabt: Weiterhin sind mehrere Systeme involviert, Entscheidungen erfolgen unkoordiniert und Aufgaben durchlaufen mehrere unzusammenhängende Schritte. Das alles soll nun KI kompensieren. Also wird der KI mehr Verantwortung erteilt: Sie übernimmt die Interpretation, Weiterleitung und Anpassung in Echtzeit, um den Prozess irgendwie zusammenzuhalten.

Dieser Ansatz kann das Erlebnis kurzfristig verbessern, ändert aber nicht die Kostenstruktur und tatsächlich steigen die Kosten häufig. Das liegt daran, dass das System nun beides erledigt: die ursprüngliche Arbeit und die ständige Interpretation, damit der Prozess wie gehabt funktioniert.

Jedes Mal, wenn das System etwas erneut bewerten oder den Arbeitsablauf rekonstruieren muss, verbraucht es zusätzliche Ressourcen. Multipliziert man das mit Tausenden oder Millionen von Interaktionen, sind die Folgen offensichtlich. Bei diesem Ansatz funktioniert KI lediglich als Ergänzung für die ständige Interpretation und ändert nichts an der zugrunde liegenden strukturellen Komplexität.

So etwas lässt sich nicht nachhaltig skalieren. Wer die Kostengleichung ändern will, muss zuerst den Prozess angehen.

Kostenkontrolle beginnt beim Design, nicht bei der Laufzeit-Optimierung

Unternehmen, die mit KI bedeutende Fortschritte erzielen, gehen das Problem anders an: Sie stellen auf ein Orchestrierungsmodell um – mit einer koordinierten, strukturierten KI, die in die Gestaltung und Ausführung von Arbeitsabläufen integriert ist.

Statt bei der Entscheidung auf KI zu setzen, nutzen diese Unternehmen die Technologie schon früher: beim Design der eigentlichen Arbeitsprozesse. In dieser Phase kann KI unglaublich leistungsstark sein, z. B. zur Analyse bestehender Prozesse, der Erkennung von Ineffizienzen oder für Vorschläge, wie sich Aufgaben über Systeme und Teams hinweg besser strukturieren lassen.

Ist diese Struktur definiert, wird die Ausführung erheblich effizienter. Das System muss nicht länger zur Laufzeit den nächsten Schritt neu „herausfinden“. Stattdessen folgt es einem vorgegebenen Pfad, bei dem KI gezielt dort eingesetzt wird, wo sie den meisten Nutzen bringt – z. B. beim Interpretieren der Absicht oder zur Handhabung von unstrukturierten Eingaben.

Diese Verlagerung vom *Reasoning zur Laufzeit* zur *KI-gestützten Designphase* ändert sowohl das Kostenprofil als auch die Verlässlichkeit der Ergebnisse. Unternehmen können so Folgendes erreichen:

- Reduzierung unnötiger Rechenleistung
- Vermeidung wiederholter Entscheidungsfindungen
- Sicherstellung, dass Aufgaben auf wiederholbare, geregelte Weise erledigt werden

Bei diesem Modell wird KI zum Teil eines Ausführungssystems – und kein Ersatz dafür.

Vorhersehbarkeit als Voraussetzung für den unternehmensweiten KI-Einsatz

Damit KI grundlegend im Unternehmen integriert wird, muss die Technologie genauso diszipliniert arbeiten wie andere Kernsysteme.

Das beginnt bei der Vorhersagbarkeit. Führungskräfte brauchen verlässliche Antworten auf drei Fragen:

- 1. Was kostet der KI-Einsatz?**
- 2. Was bringt uns das?**
- 3. Wie funktioniert die KI in der Praxis?**

Ohne Klarheit in diesen Punkten wird die Skalierung riskant.

Wie eine aktuelle EY-Studie zeigt, fehlt bei über der Hälfte der KI-Initiativen eine formale Aufsicht – was sich bei zunehmender Skalierung in Kostenüberschreitungen, Compliance-Lücken und unzuverlässigen Ergebnissen niederschlägt.



An diesem Punkt kommt ein strukturierter KI-Ansatz zum Tragen. Wenn KI innerhalb klar definierter Workflows arbeitet – unterstützt durch Orchestrierung und Governance –, geschieht etwas Entscheidendes:

- Die Ergebnisse werden zuverlässiger.
- Die Kosten sind besser planbar und sind zunehmend an Entscheidungen, Transaktionen und belegbare Geschäftsergebnisse geknüpft.
- Operative Abläufe lassen sich leichter managen.

Und was am wichtigsten ist: Diese Vorhersagbarkeit gilt nicht nur für die Token-Nutzung, sondern für alle Gesamtbetriebskosten – vom laufenden Betrieb und der Governance über die Compliance bis hin zu Investitionen in die kontinuierliche Weiterentwicklung der Systeme. Hierbei werden die Gesamtbetriebskosten deutlicher, weil neben den Token-Kosten auch die Kosten für den Betrieb, die Kontrolle, die Pflege und die Skalierung von KI im Laufe der Zeit berücksichtigt werden.

Statt KI als eine Art Tool ohne Grenzen zu betrachten, müssen Unternehmen die Technologie als kontrollierten, messbaren Teil von Arbeitsabläufen behandeln. Dann gelingt der Schritt von der reinen Verbrauchserfassung hin zur Verfolgung der Ergebnisse. Und das ist ein wichtiger Schritt, wenn man KI wirtschaftlich einsetzen und davon profitieren will.

⁴ EY, „Autonomous AI adoption surges as oversight falls behind“ (2026)

Erfolgreiche Unternehmen werden ihre Sicht auf KI ändern

In vielen Unternehmen herrscht weiterhin die Ansicht, dass man sich zwischen einem schnellen oder kontrollierten KI-Einsatz entscheiden muss. Tatsächlich setzt nicht die Geschwindigkeit die Grenzen, sondern die Architektur.

Denn die schnelle Einführung von KI allein garantiert nicht, dass das Unternehmen vom gesamten Potenzial der Technologie profitiert.

Unternehmen, die dies richtig machen, überdenken grundlegend, wie ihr Geschäft funktioniert. Konkret bedeutet das, dass der Schwerpunkt auf anderen Prioritäten liegt:

Schwerpunkt auf Systeme statt Tools

Entscheidend ist nicht, was KI für sich genommen leisten kann, sondern wie gut sie sich in unternehmensweite Arbeitsabläufe integrieren lässt.

Neugestaltung vor der Automatisierung

Wird KI auf einen nicht funktionierenden Prozess angewendet, verbessert das meistens gar nichts. Wird dagegen der Prozess zuerst neu gestaltet, entsteht die Grundlage für richtungsweisende Veränderungen.

Ergebnisse messen, nicht die Aktivität

Erfolg sollte an Geschäftsergebnisse geknüpft sein, wie abgeschlossene Aufgaben, verbesserte Erlebnisse oder weniger Reibungsverluste – nicht an die Frage, wie oft KI genutzt wird.

KI als operative Infrastruktur behandeln

Wenn KI stärker in Kernprozesse eingebettet wird, muss die Technologie die gleichen Zuverlässigkeits- und Governance-Standards wie jedes andere Unternehmenssystem erfüllen.

Im Vordergrund steht der disziplinierte Einsatz von neuen Technologien, statt unüberlegt auf jeden Zug aufzuspringen.

Von Experimenten zur praktischen Umsetzung

Wir stehen bei Unternehmens-KI an einem Wendepunkt. In der Frühphase haben wir gesehen, was möglich ist. Die nächste Phase wird definieren, was praktisch ist und was sich skalieren lässt.

Wenn Unternehmen immer mehr KI einsetzen, besteht die Herausforderung nicht mehr darin, Insights zu gewinnen oder einzelne Aufgaben zu automatisieren. Es geht dann darum, dass diese Funktionen stabil, effizient und auf Geschäftsergebnisse ausgerichtet genutzt werden können.

Das ist der Unterschied zwischen vereinzelteten Erfolgen und Vorteilen, von denen das gesamte Unternehmen profitiert.

Unternehmen, die diese Lücke schließen, werden schneller handeln, effizienter funktionieren und sich besser auf Veränderungen einstellen können.

Aber das lässt sich nicht erreichen, indem man KI einfach unkoordiniert erweitert.

Entscheidend ist, dass das die richtige Grundlage für den unternehmensweiten KI-Einsatz geschaffen wird.

Fazit

KI verändert schon heute die Möglichkeiten von Unternehmen.

Aber was jetzt zählt, ist, was die Technologie skaliert in der Praxis leistet – über komplexe Systeme hinweg und innerhalb von Kosten- und Kontrollvorgaben.

Die meisten versteckten KI-Kosten verursacht nicht das Modell, sondern die Strukturierung, Ausführung und Wiederholung von Aufgaben im Laufe der Zeit.

Den Unternehmen, die dieses Problem lösen, geht es nicht nur um Kostensenkung: Sie wollen ein System schaffen, das für ihr Geschäft wirklich funktioniert.

Und in der nächsten Ära der Enterprise-KI werden diese Unternehmen die Vorreiter sein – weil sie eine KI haben, die sie kontrollieren, messen und als Teil des Unternehmens betreiben können.



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)**