1. KOSTEN, PRODUKTIVITÄT UND DIE NEUE ÖKONOMIE VON AGENTURARBEIT

KI senkt Kosten nicht. Sie verschiebt sie.

Warum KI nicht kostenlos ist

Die Wahrnehmung von KI-Kosten hat drei Phasen durchlaufen. In der ersten Phase wirkte KI spektakulär und weitgehend kostenlos. Große Modelle waren im Browser erreichbar, einzelne Abonnements kosteten weniger als viele etablierte Softwarelizenzen. Kostenlose Einstiegsangebote und Probezugänge machten die Nutzung leicht. In vielen Agenturen und Marketingabteilungen wurde experimentiert, ohne dass die Kostenfrage ernsthaft gestellt wurde.

Diese Wahrnehmung knüpft an eine durch digitale Dienste geprägte Erwartung an. Viele Märkte wurden durch das Internet daran gewöhnt, dass die Nutzung digitaler Tools scheinbar kostenlos oder nahezu kostenlos zu haben sei. Suchen, soziale Netzwerke, E-Mail-Dienste und Plattformzugänge wurden selten unmittelbar bezahlt. Monetarisiert wurde später, indirekt oder an ganz anderen Stellen: über Werbung, Daten, Abhängigkeit, Firmen- und Großkundentarife oder geschlossene Ökosysteme. Diese Erfahrung prägt auch den Blick auf KI. Was im Browser verfügbar ist, wirkt wie ein weiterer digitaler Dienst mit geringen Grenzkosten. Doch diese Schlussfolgerung ist falsch. KI ist nicht nur Oberfläche. Jede Nutzung greift auf Modelle, Rechenleistung, Speicher, Energie, Schnittstellen und Betriebsinfrastruktur zu.

In der zweiten Phase wurde KI wie andere Software eingeführt: ChatGPT Plus für einzelne Teams, Midjourney-Konten oder Firefly-Guthaben in der Kreation für die Bildproduktion, Copilot-Lizenzen für Office-Umgebungen. Die Kosten waren sichtbar, aber kalkulierbar. Sie wurden meist wie normale Toolkosten behandelt.

In der dritten Phase, die derzeit am Anfang steht, wird deutlich, dass für KI andere wirtschaftliche Regeln gelten. Sie ist mehr als Software. Sie wird Infrastruktur. Und Infrastruktur wird bepreist, sobald sie dauerhaft in Arbeitsprozessen, Angeboten und Kundensystemen genutzt wird.

Die Anfangsphase war keine stabile Zielökonomie. Sie war Teil einer Marktdurchdringung. Anbieter mussten Nutzung aufbauen, Gewohnheiten schaffen, Ökosysteme besetzen und Abhängigkeiten etablieren. Sobald KI nicht mehr ausprobiert, sondern in tägliche Arbeit, Kundenprozesse und Produktangebote integriert wird, verändert sich auch die Preislogik: aus Zugang wird Verbrauch, aus Nutzung wird Betrieb, aus dem einst günstigen Tool wird ein reeller Kostenfaktor in der Wertschöpfung.

Wie unterschiedlich dieselbe Entwicklung bewertet wird, zeigt der Blick auf die Anbieter- und die Anwenderseite: Was für Infrastrukturunternehmen ein wachsendes Geschäft bedeutet, führt bei Unternehmen und Agenturen zunächst zu steigenden Kosten. Diese Verschiebung zeigt sich besonders deutlich in der Sprache der Infrastruktur- und Plattformanbieter. NVIDIA-CEO Jensen Huang beschreibt Compute inzwischen als direkten Weg zu Intelligenz, Umsatz und Wachstum. Aus dieser Perspektive werden Rechenzentren zu „AI Factories“, Tokens zu produzierten Einheiten und „Tokens per Watt“ zu einer zentralen Effizienzkennzahl. Für NVIDIA ist diese Perspektive konsequent: Mehr Modellnutzung bedeutet mehr Infrastrukturumsatz. Für Unternehmen und Agenturen gilt jedoch nicht „Compute = Revenue“, sondern zunächst: **Compute = Kostenbasis**. Mehr Tokens erzeugen nicht automatisch mehr Wirkung. Sie erzeugen variable Kosten, Abhängigkeit und Steuerungsbedarf. Entscheidend wird deshalb nicht sein, wie viel KI genutzt wird, sondern ob zusätzliche Modellnutzung nachweisbar zu besseren Entscheidungen, besserer Kommunikation oder höherem Marktwert führt.

Das zeigt sich bereits in den Preisarchitekturen der großen Anbieter. OpenAI rechnet nach Eingabe-, zwischengespeicherten und Ausgabe-Einheiten ab und bepreist zusätzliche Funktionen wie Dateisuche oder temporäre Arbeitsumgebungen separat. Anthropic differenziert nach Modellklassen und der Zwischenspeicherung von Eingaben. Google Gemini und Vertex AI verbinden Modellnutzung mit Speicher-, Laufzeit- und Infrastrukturkosten. Adobe Firefly strukturiert die Bildproduktion über Nutzungsguthaben und Tarifstufen. Das ist keine Tarifverwirrung, sondern Marktreifung. Was als Zugang begann, wird zur differenzierten Verbrauchs- und Infrastrukturökonomie.

Für Agenturen und Unternehmen entsteht auf diese Weise ein Kalkulationsproblem. Ein KI-gestützter Arbeitstag kann heute aus Softwarelizenzen, Verbrauch über Schnittstellen, Bildguthaben, Suchabfragen, Speicher, Rechenzeit und menschlicher Qualitätssicherung bestehen. Einzelne Workflows generieren oft mehrere die-

ser Kostenströme, die nur selten an einer Stelle zusammenlaufen. Sie erscheinen in Kreditkartenabrechnungen, IT-Budgets, Projektstunden, Cloud-Konten oder gar nicht, etwa weil Mitarbeitende private Tools nutzen. Hier beginnt die neue Kostenfrage: Wird die gesamte KI-gestützte Leistung wirtschaftlich verstanden und bleibt sie steuerbar?

Hinzu kommt: Der Markt kennt kein einheitliches Preismodell. Westliche Anbieter wie OpenAI, Anthropic, Google, Microsoft oder Adobe monetarisieren Zugang, Nutzung, Firmen- und Großkundenfunktionen sowie Infrastruktur zunehmend konsequent. Gleichzeitig setzen große chinesische Anbieter wie Qwen, DeepSeek, Z.ai, Moonshot oder MiniMax stark auf offene Modelle, breite Verfügbarkeit und Ökosystemstrategien. Verdient wird dort insbesondere an der Ausführung in der Cloud, an Integration, Infrastruktur oder dem Folgegeschäft.

Diese Gleichzeitigkeit macht die Kostenbasis von KI instabil. Auf der einen Seite steigen Preise, werden Funktionen ausdifferenziert und Großkundenmodelle ausgebaut. Auf der anderen Seite drücken offene Modelle und alternative Anbieter auf den Markt. Für Agenturen und Unternehmen ist daher wichtig: KI ist nicht grundsätzlich günstig oder teuer. Die Frage ist, wie abhängig ein Workflow, ein Service oder ein Angebot von einer bestimmten Preis-, Modell- und Anbieterstruktur ist.

Das betrifft zusätzlich zur Content-Produktion Research, Datenanalyse, Automatisierung, Programmierung, Monitoring, Übersetzung, Reporting, Prototyping und agentische Workflows. Überall dort, wo KI dauerhaft in Leistungen eingebaut wird, entsteht eine Kostenbasis, die sich jederzeit verändern kann. Wenn Anbieter Preise, Limits, Schnittstellen, Nutzungsrechte oder Abrechnungsmodelle ändern, kippt mehr als einzelne Rechnungen. Es können ganze Services, Toolanbieter, Angebotsmodelle und Workflows unter Druck geraten. Strategisch erleben wir deshalb den Übergang von einer Lizenzökonomie zu einer Verbrauchsökonomie. Dieser Übergang ist in vielen Controllingprozessen noch nicht angekommen.

KI ist nicht kostenlos, und sie bleibt auch nicht automatisch günstig, nur weil die Nutzung an der Oberfläche einfach wirkt. Ein Prompt ist nicht nur eine Eingabe. Er kann eine Produktionskette auslösen. Ein Schnittstellenaufruf ist nicht nur ein technischer Vorgang. Er kann externe Kosten erzeugen. Ein KI-gestützter Programmierassistent spart Entwicklungszeit, erzeugt aber neue Prüf-, Sicherheits- und Betriebskosten. Eine Automatisierung ist kein Selbstläufer. Sie muss überwacht, gepflegt und verantwortet werden. Es entstehen Kosten, die in vielen Kalkulationen noch fehlen.

KI-gestützte Arbeit muss anders kalkuliert werden als klassische Software-Nutzung oder klassische Arbeitszeit. Wer nur heutige Toolpreise betrachtet, kalkuliert zu kurz. Wer KI-Leistungen professionell anbietet oder einkauft, muss die Kostenarchitektur dahinter verstehen: Lizenzen, Verbrauch, Rechenleistung, Daten, Qualität, Governance, Betrieb und Anbieterabhängigkeit. Erst dann lässt sich seriös beantworten, was KI wirklich kostet und wer die anfallenden Kosten zu tragen hat.

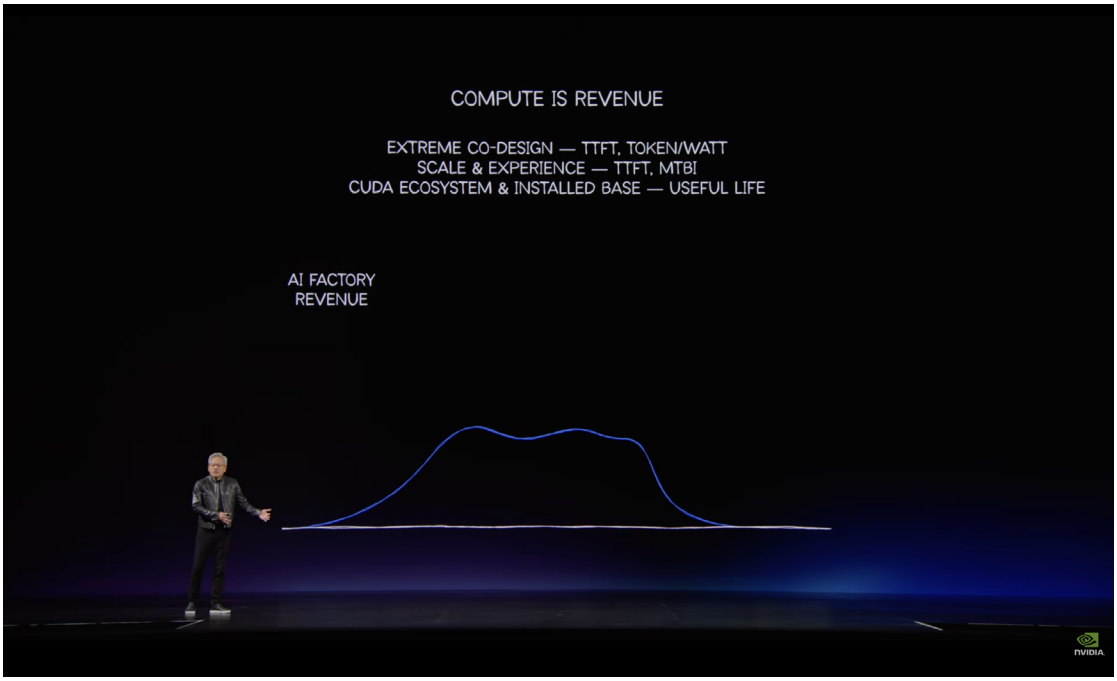
Kostenmodelle der KI-Arbeit

Die klassische Agenturökonomie kennt eine dominante Kostenbasis: Gehälter. Menschen werden eingestellt, ausgelastet und über Stundensätze, Tagessätze, Pauschalen oder Retainer kalkuliert. Darauf beruhen Auslastungsplanung, Margenkalkulation und große Teile der Agentursteuerung. Dieses Prinzip verschwindet durch KI nicht, aber sie reicht nicht mehr aus, um KI-gestützte Arbeit wirtschaftlich zu verstehen.

Neben Gehalt und klassischer Toolnutzung treten Kosten, die in Projekten, Workflows und Kundenbeziehungen gleichzeitig wirken: **Lizenz, Token und Compute**. Sie stehen für drei unterschiedliche Fragen: Wer hält den Zugang? Wer steuert den Verbrauch? Wer verantwortet den Betrieb?

Denn darin liegt der Unterschied zu klassischen Kosten. Eine Agentur rechnet in der Regel nicht jede Adobe-Lizenz, jeden Projektmanagement-Account oder jedes Studiolicht einzeln ab und reicht die entstandenen Kosten an den Kunden weiter. Solche Kosten sind meist Teil einer internen Kalkulation, der Gemeinkosten und des Geschäftsmodells. Bei KI wird diese Logik infrage gestellt, denn KI-Kosten können variabel, nutzungs-, workflow-, kunden- und sogar plattformabhängig sein. Sie entstehen durch den Zugang zu einem Tool und außerdem durch die Nutzung von Datenvolumen, Schnittstellen und Rechenleistung sowie durch Qualitätssicherung und laufenden Betrieb.

Das erste Kostenmodell ist die **Lizenz**. Sie folgt dem vertrauten Modell der Softwarenutzung: ein



Quelle: NVIDIA, „GTC Taipei 2026 Keynote | Full Replay", ↗ Vortrag von Jensen Huang, ab Min. 35:25, abgerufen am 08.09.2026.

fixer Betrag pro Nutzer:in, Team, Funktionspaket oder Monat. Diese Logik ist planbar und passt in bestehende Budgets. Ihre Schwäche liegt darin, dass sie wenig über die tatsächliche Nutzung, Produktivität und Verantwortung aussagt. Eine Lizenz kann ungenutzt bleiben. Sie kann aber auch der Zugangspunkt zu sensiblen Daten, geschützten Kundeninformationen oder markenrelevanten Workflows sein. Dann ist sie zusätzlich zum Kostenpunkt eine Governance-Frage.

Das zweite Kostenmodell basiert auf **Tokens**. Gemeint sind die variablen Nutzungskosten, die bei KI-Systemen durch Eingaben, Ausgaben, Suchvorgänge, Bildgenerierungen, Dokumentenanalysen, Schnittstellenaufrufe oder automatisierte Prozessläufe entstehen. Diese Logik ist für Agenturen und Unternehmen ungewohnt, weil sie Volumen, Nutzungstiefe und Iterationen kalkulierbar machen muss, bevor diese vollständig bekannt sind. Ein einzelner Research-Workflow mit zehn Dokumenten entspricht einer anderen Kostenrealität als einer mit tausend Dokumenten. Ein Monitoring-Prozess, der täglich läuft, ist etwas anderes als eine einmalige Analyse. Tokens stehen damit für eine technische Abrechnungseinheit und für die Frage, wer den Verbrauch auslöst, steuert oder sogar begrenzt.

Das dritte Kostenmodell betrifft **Compute**. Gemeint sind Infrastruktur- und Betriebskosten: Rechenleistung, Speicher, Laufzeit, Datenpflege, Monitoring, Qualitätssicherung, Schnittstellenbetrieb, Support und laufende Weiterentwicklung. Diese Logik wird besonders relevant, wenn KI nicht mehr nur als einzelnes Werkzeug genutzt wird, sondern in Workflows, Kundensysteme oder wiederkehrende Leistungen integriert wird. Dann geht es mehr darum, ob ein System zuverlässig läuft, als darum, ob ein Tool genutzt wird.

Diese drei Kostenmodelle überlagern sich. Ein KI-gestützter Workflow kann gleichzeitig aus Nutzerlizenzen, variablen Nutzungskosten, Rechenleistung, Speicher, Datenpflege und menschlicher Prüfung bestehen. Hinzu kommt die klassische Personalkostenlogik: Menschen briefen, entscheiden, prüfen, verdichten, verantworten und übersetzen Ergebnisse in den Kontext von Marke, Markt und Organisation. Gerade diese menschliche Steuerung bleibt kalkulationsrelevant, auch wenn einzelne Ausführungsschritte schneller werden.

Kalkulation wird zur Verteilungsfrage. Welche Kosten sind internes Betriebsmittel der Agentur? Welche entstehen durch einen konkreten Kundenauftrag? Welche gehören in die Infrastruktur des Unternehmens? Welche werden gemeinsam betrieben? Welche müssen als laufende Systemleistung bepreist werden? Und welche Verantwortung ist mit welcher Kostenlogik verbunden?

Diese Fragen werden umso wichtiger, je stärker werbetreibende Unternehmen Leistungen internalisieren wollen, die bisher extern waren.

Die drei Kostenlogiken der KI

	LIZENZ Seat-basiert	**VERBRAUCH** Token-basiert	**COMPUTE** Infrastruktur-basiert
Abrechnung	Fixer Preis je Nutzer:in und Monat	Pro Einheit Nutzung (Tokens, Tool-Calls)	Pro Recheneinheit (vCPU, GPU, Speicher)
Beispiele	ChatGPT Plus, Copilot, Firefly-Plan, Midjourney	OpenAI API, Claude API, Gemini API	Qwen, DeepSeek, Z.ai (Open Source, self-hosted)
Kostentreiber	Anzahl aktiver Seats	Nutzungsintensität und Kontextgröße	Auslastung der Infrastruktur
Ineffizienzfalle	Ungenutzte Seats	Ungesteuerte Dauer-abfragen	Unterausgelastete Kapazität
Budgetzuordnung	IT / Lizenzen	Projekt / Cloud	Infrastruktur

Drei nebeneinander existierende Abrechnungsmodelle mit je eigener Ökonomie.

Inhousing verlagert Aufgaben in die eigene Organisation. Es klärt aber nicht automatisch, auf welcher Infrastruktur diese Aufgaben laufen, wer die Nutzungsbedingungen akzeptiert, welche Daten verarbeitet werden, wer Qualität prüft, wer den Verbrauch steuert und wer Verantwortung übernimmt, wenn sich Preise, Modelle oder rechtliche Rahmenbedingungen ändern. Wenn Unternehmen mehr Research, Content, Analyse, Automatisierung oder Programmierung intern abdecken wollen, müssen sie deshalb zusätzlich zu Rollen und Fähigkeiten auch die Kosten-, Lizenz- und Verantwortungslogik dahinter klären.

Gerade größere Unternehmen haben Procurement, Datenschutz, IT-Sicherheit, Legal, Compliance, Markenführung und oft auch betriebliche Mitbestimmung im System. Dort lassen sich neue Nutzungsbedingungen nicht einfach per Klick akzeptieren. Wenn zudem regulierte oder vertrauliche Informationen, Kundendaten, Markenwissen oder automatisierte Prozesse betroffen sind, entstehen neue Schnittstellen und Prozessketten zwischen internen Teams, externen Partnern und Technologieanbietern. Der Wunsch nach Inhousing führt deshalb nicht automatisch zu mehr Eigenständigkeit, Effizienz oder Effektivität.

Für Agenturen entstehen daraus neue Rollen. Sie können KI-gestützte Leistungen entwickeln und ebenfalls orchestrieren, betreiben und absichern. Das kann als Beratung und Enablement geschehen, als Betrieb eines gemeinsamen Workflows, als kuratierter Zugang zu Tools und Modellen oder perspektivisch auch als Plattform- oder Vermittlungsmodell. Dreh- und Angelpunkt ist dabei, wie die zugrunde liegenden Kosten, Lizenzen und Verantwortlichkeiten sichtbar und vertraglich geregelt sind.

Die Lizenzfrage wird damit zur Governance-Frage: Wer hält den Zugang? Wer akzeptiert die Nutzungsbedingungen? Wer darf welche Daten eingeben? Wem gehören Arbeitsumgebungen, Prompts, Templates, Markenwissen und Prozesslogiken? Wer trägt die Verantwortung, wenn ein Anbieter seine Preise, Funktionen, Schnittstellen oder Nutzungsbedingungen ändert? Eine KI-Lizenz ist mehr als Beschaffung. Sie definiert Datenwege, Nutzungsgrenzen, Abhängigkeiten und Verantwortlichkeiten.

Die strategische Aufgabe besteht deshalb darin, KI-Kosten als gemeinsame Kosten- und Verantwortungsarchitektur zu behandeln. Wer diese Architektur früh versteht, kann realistischer kalkulieren, Angebote sauberer formulieren, Margen schützen und Kundenbeziehungen belastbarer gestalten. Wer sie ignoriert, riskiert falsche Preise, unklare Zuständigkeiten und Leistungen, deren wirtschaftliche Grundlage später kippt.

Produktivität ist nicht gleich Einsparung

Die Befunde zur KI-Produktivität sind inzwischen deutlich. Bei realisierten Einsparungen ist das Bild weniger eindeutig. Zwischen Produktivität und Einsparung liegt eine Lücke, die für die Agenturökonomie und für die Einkaufslogik von Unternehmen strategisch zentral ist.

Auf der Produktivitätsseite liefert die Feldstudie von Brynjolfsson, Li und Raymond eine der belastbarsten Referenzen. Im Kundenservice misst sie nach Einführung generativer KI-Assistenten einen durchschnittlichen Produktivitätszuwachs von 14 %. Der Durchschnittswert erzählt jedoch nur die halbe Geschichte. Besonders stark profitieren Berufseinsteiger:innen und weniger erfahrene Mitarbeitende. Bei ihnen lag der Produktivitätszuwachs bei bis zu 34 %. Bei erfahrenen Hochleister:innen waren die Effekte deutlich geringer, in einzelnen Teilaufgaben sogar negativ.[5] Die OECD berichtet in einer Metaanalyse aus dem Jahr 2025 ähnliche Größenordnungen in anderen Tätigkeitsfeldern, mit Bandbreiten zwischen 5 und über 25 % je nach Aufgabe und Kontext.[6]

Auf der Einsparungsseite zeigt sich ein anderes Bild. McKinsey berichtet in der State of AI Studie 2025, dass mehr als 80 % der befragten Unternehmen auf Gesamtebene noch keinen materiellen EBIT-Effekt durch generative KI messen konnten, obwohl viele von ihnen gleichzeitig Kostensenkungen in einzelnen Funktionen beobachteten.3 Während funktionale Produktivitätsgewinne also bereits sichtbar werden, entfalten sich aggregierte wirtschaftliche Effekte deutlich langsamer – zumindest nicht in dem Tempo, das mit der öffentlichen Erwartung Schritt hält. Produktivität ist zunächst ein Effekt auf Aufgabenebene, Einsparung hingegen ist ein Effekt auf Organisations-, Budget- und Geschäftsmodellebene. Dazwischen liegen Prozesse, Erwartungen, Verträge, Steuerung und Zeit. KI spart also nicht automatisch Geld. Sie schafft Spielraum. Was aus diesem Spielraum wird, ist eine Management- und Verhandlungsfrage.

Vier Filter halten Produktivität häufig davon ab, direkt als Einsparung sichtbar zu werden.

Der erste Filter ist Absorption. Gewonnene Zeit wird häufig in zusätzliche Leistung umgeleitet und nicht als Einsparung realisiert: Dasselbe Briefing bekommt drei statt einen Moodboard-Vorschlag. Derselbe Text wird in drei Tonalitäten statt in einer erarbeitet. Dieselbe Recherche wird breiter angelegt, weil es möglich ist. Die Qualität kann steigen, ohne dass die Kosten im gleichen Maß sinken.

Der zweite Filter ist Erwartungsinflation. Kunden erwarten schneller, mehr und variantenreicher, was vorher länger dauerte. Was produktiver wird, wird teilweise zur neuen Mindestanforderung, bevor es als zusätzlicher Wert bepreist werden kann. Eine schnelle Lieferung ist in vielen Kundenbeziehungen kein Verkaufsargument mehr, sondern Voraussetzung.

Der dritte Filter sind komplementäre Investitionen. Produktivitätsgewinne auf Aufgabenebene bleiben an der Oberfläche, solange Workflows, Prozesse, Zuständigkeiten und Schnittstellen unverändert bleiben. Wirtschaftliche Effekte entstehen dort, wo nach der Einführung von KI die Arbeitsweise entlang der neuen Möglichkeiten umgebaut wird. Ohne diesen Umbau bleibt die Geschwindigkeit an der einzelnen Tastatur und kommt im Jahresabschluss nicht an.

Der vierte Filter ist Messträgheit. Produktivität lässt sich in einzelnen Aufgaben schnell beobachten. Einsparung braucht Budgetzyklen, Kostenstellen, Vertragslaufzeiten und Managemententscheidungen, um sichtbar zu werden. Ein Gewinn in der Produktion landet erst dann im Ergebnis, wenn entsprechende Personal-, Sach- oder Fremdkosten tatsächlich reduziert, umgeschichtet oder in zusätzliche Leistung übersetzt werden.

Für Agenturen kommt eine besondere Spannung hinzu. Das klassische Agenturmodell verdient häufiger an verkaufter Arbeitszeit als an produzierter Einheit oder erzeugter Wirkung. Wenn ein Text schneller geschrieben wird, ein Moodboard schneller entsteht oder eine Kampagnenoption schneller entwickelt ist, reduziert sich im Stundenmodell zuerst die abrechenbare Zeit. Produktivitätsgewinne ohne Anpassung von Kalkulation, Scope oder Geschäftsmodell sind in dieser Logik kein automatischer Gewinn. Sie können sogar Umsatz und Marge unter Druck setzen.
Für Unternehmen ist die Gegenbewegung ebenso wichtig. Schnellere Arbeit bedeutet nicht automatisch niedrigere Gesamtkosten. Wenn mit KI mehr Varianten, mehr Geschwindigkeit, mehr Qualität, mehr Sicherheit und mehr Nachvollziehbarkeit erwartet werden, wird Produktivität in zusätzliche Leistung umgewandelt. Deshalb reicht es nicht, KI-Effizienz nur als Preissenkungsargument zu lesen.

5 Brynjolfsson, E., Li, D. & Raymond, L. R. (April 2023). Generative AI at Work. NBER Working Paper No. 31161. National Bureau of Economic Research. ↗ Abgerufen am 08.09.2026

6 Calvino, F., Reijerink, J. & Samek, L. (2025). The effects of generative AI on productivity, innovation and entrepreneurship. OECD Artificial Intelligence Papers, No. 39. OECD Publishing, Paris. ↗ Abgerufen am 08.09.2026

Was nicht als Einsparung realisiert wird, verschwindet nicht. Es taucht an anderer Stelle wieder auf: in höheren Qualitätserwartungen, in mehr Projektvolumen, in schnelleren Zyklen, in neuen Prüfprozessen oder in Leistungen, die die Organisation vorher nicht erbracht hat. Die entscheidende Frage lautet deshalb: Wohin fließt der Produktivitätsgewinn? Er kann als Einsparung beim Kunden landen. Er kann als Marge bei der Agentur bleiben. Er kann in bessere Qualität, mehr Varianten, schnellere Zyklen oder neue Leistungen fließen. Wirtschaftlich relevant wird KI erst dort, wo diese Verteilung bewusst gestaltet wird.

Wo KI Kosten senkt und wo sie neue Kosten erzeugt

Wo KI Stückkosten senkt, ist relativ klar beschreibbar. Wo sie neue Grundkosten erzeugt, ist weniger offensichtlich. Darin liegt die ökonomische Verschiebung: Einzelne Einheiten können schneller und günstiger entstehen, aber daraus folgt noch keine Einsparung im Gesamtprozess. Eine Rohfassung, eine Übersetzung, eine Bildvariante oder eine erste Recherche wird erst dann wirtschaftlich günstiger, wenn klar ist, wie viele Varianten entstehen sollen, wann ein Ergebnis gut genug ist und wer entscheidet, dass der Prozess endet.

Ohne diese Steuerung wird aus sinkenden Stückkosten keine Einsparung, sondern Überproduktion. Aus einer Rohfassung werden zehn, aus zehn Varianten hundert, aus einer ersten Recherche ein offener Suchprozess. Die Kosten verschwinden also nicht. Sie verlagern sich von der Erstellung einzelner Einheiten hin zu Auswahl, Prüfung, Abstimmung und Entscheidung. KI macht Arbeit schneller. Wirtschaftlich günstiger wird diese erst, wenn Prozesse bewusst begrenzt, qualitätsgesichert und in klare Abläufe und Verantwortlichkeiten übersetzt werden. Die sichtbarsten Einsparpotenziale liegen deshalb in standardisierbaren Arbeitsschritten: zum Beispiel Rohtexte, erste Entwürfe, einfache Recherchen, Standardübersetzungen, Bildvarianten, Datenaufbereitung und modulare Produktionsschritte. Dort kann KI den Aufwand pro Einheit deutlich reduzieren. Was früher ein eigenes kleines Projekt war, kann heute Teil eines stärker automatisierten Ablaufs werden. Voraussetzung ist aber, dass der Ablauf vorher definiert ist: Input, Qualitätsmaßstab, Variantenlogik, Freigabe und Endpunkt.

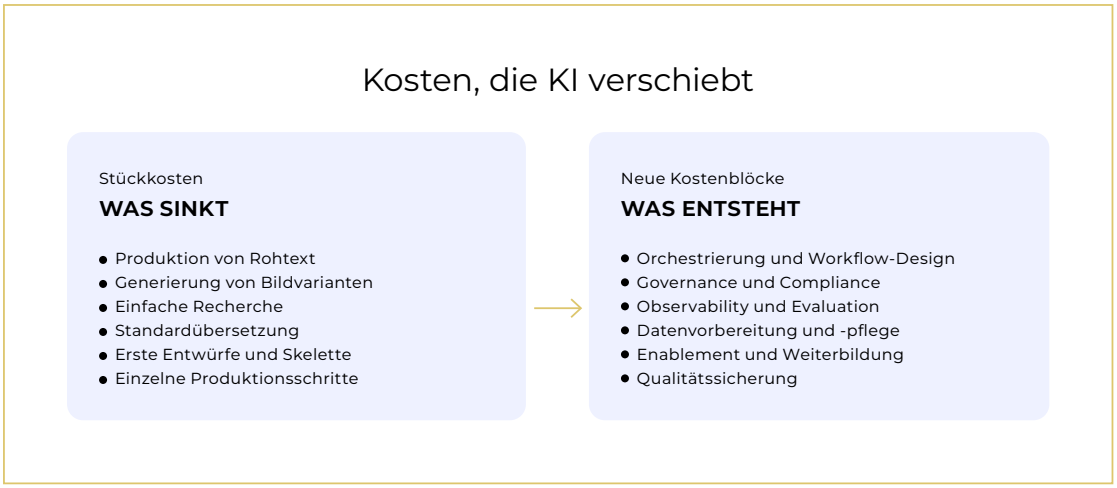
Weniger eindeutig ist der Effekt bei kreativen und strategischen Kernleistungen. Die Idee, die Perspektive, das Urteil, die strategische Einordnung, das Gespräch mit den Kunden, das Briefing, das Einfangen einer Markenwahrheit – diese Leistungen werden nicht automatisch günstiger, nur weil einzelne Vorarbeiten schneller entstehen. Im Gegenteil: Je schneller mehr Material erzeugt wird, desto wichtiger wird die Arbeit, die dieses Material sortiert, verdichtet und verantwortet. Die Produktionsgeschwindigkeit erhöht den Bedarf an kuratorischer und strategischer Leistung. So verschiebt sich die Kostenlogik. Die sichtbaren Kosten einzelner Outputs können sinken, während neue Grundkosten für Steuerung, Qualität, Daten, Regeln und Betrieb entstehen. Professionelle KI-Leistung besteht mehr als in Geschwindigkeit darin, jene Bedingungen zu schaffen, unter denen Ergebnisse verlässlich, markenkonform, rechtssicher und nutzbar werden.

Diese neue Grundkostenmodell berührt sechs Bereiche.

Ablaufsteuerung. Wer KI produktiv einsetzen will, muss festlegen, welches System oder Modell für welche Aufgabe genutzt wird, in welcher Reihenfolge gearbeitet wird, wo menschliche Prüfung nötig ist und welche Qualitätsstandards und Regeln gelten. Ein KI-gestützter Workflow besteht nicht aus einem Knopfdruck oder einem Prompt, sondern aus Entscheidungen: Welche Daten werden verwendet, welche Varianten erzeugt? Was wird verworfen? Wann wird eskaliert? Was darf automatisiert werden und was nicht?

Freigaben und rechtliche Absicherung. Je stärker KI in Kundenarbeit eingesetzt wird, desto relevanter werden Fragen nach Daten, Rechten, Kennzeichnung, Vertraulichkeit und Verantwortung. Welche Daten dürfen in welches System? Welche Inhalte brauchen besondere Prüfung? Welche Nutzung ist vertraglich erlaubt und abgesichert? Welche Ergebnisse müssen dokumentiert oder gekennzeichnet werden, und wem gehören sie? Diese Arbeit begleitet den gesamten Prozess und entsteht nicht erst am Ende.

Qualitätssicherung. Gerade weil KI schnell viele plausible Ergebnisse erzeugt, steigt die Bedeutung der Prüfung. Qualitätssicherung ist nicht mehr nur die letzte Korrekturschleife vor der Veröffentlichung, sondern eine parallel laufende Aufgabe. Ergebnisse müssen geprüft, verglichen, bewertet und in den jeweiligen Marken- und Nut-



Die Ökonomie der Agenturarbeit wird umgebaut, nicht verkleinert.

zungskontext übersetzt werden. Qualität entsteht durch bessere Auswahl, nicht durch mehr Output.

Datenpflege und Wissensgrundlagen. KI-gestützte Arbeit wird besser, wenn sie auf sauberen Referenzen, gepflegten Markeninformationen, klaren Tonalitäten, konkreten Beispielen und Produktdaten sowie auf Kundenvokabularen aufsetzt. Diese Grundlagen müssen aktiv aufgebaut, aktualisiert und geschützt werden. Andernfalls verschlechtert sich die Qualität der Ergebnisse oft schleichend, bevor dies in den Projekten sichtbar wird.

Überwachung laufender Systeme. Wer KI in wiederkehrende Prozesse einbaut, muss beobachten, ob Ergebnisse stabil bleiben oder ob sich Fehler einschleichen, ob Systeme anders reagieren als erwartet oder ob neue Modellversionen die Ergebnisqualität verändern. Das gilt besonders für automatisierte Abläufe und agentische Systeme. Was einmal eingerichtet wurde, bleibt nicht automatisch zuverlässig.

Enablement und Weiterbildung. Neue Modelle, neue Funktionen und neue Anwendungsformen verändern laufend, was sinnvoll möglich ist. Schulung ist deshalb kein einmaliges Einführungsprojekt, sondern Teil des Betriebs. Teams müssen lernen, wie sie KI nutzen, wo sie ihr misstrauen müssen, wie sie Ergebnisse bewerten und wie sie Verantwortung übernehmen.

Diese Tätigkeiten sind nicht alle neu, aber neu ist ihre Rolle im System. Sie werden von begleitenden Tätigkeiten zu Voraussetzungen professioneller KI-Leistung. Wer sie nicht sichtbar macht, unterschätzt den tatsächlichen Aufwand. Wer sie als Nebenkosten behandelt, finanziert sie zu knapp. Und wer sie nicht in Angebote und Kundenbeziehungen übersetzt, trägt sie am Ende selbst.

Für Unternehmen ist das gleichermaßen relevant wie für Agenturen. Wer KI-gestützte Leistungen einkauft, kauft die Steuerung der Bedingungen, unter denen diese Produktion sinnvoll, sicher und wirksam wird. Wenn Unternehmen nur den günstigeren Output sehen, aber Steuerung, Qualität, Datenpflege, Freigaben und Betrieb nicht berücksichtigen, entsteht Preisdruck an der falschen Stelle.

Die entscheidende Frage lautet daher nicht nur, wo KI Kosten senkt, sondern auch, welche neuen Kosten entstehen, damit die sinkenden Stückkosten überhaupt realisiert werden können. KI kann einzelne Outputs günstiger machen. Professionelle KI-Leistung wird aber erst dort wirtschaftlich tragfähig, wo Geschwindigkeit begrenzt, Qualität gesichert und Verantwortung organisiert wird. Und dort entsteht die neue Kostenrealität.

Standardisierung, Skalierung und neue Stückkosten

Stückkosten sind in der Agentur- und Produktionswelt kein neues Konzept. Wiederholbare Leistungen wurden schon immer kalkuliert: bei Adaptionen, Übersetzungen, Reinzeichnungen, Medienproduktion, Bildbearbeitung, Formatvarianten oder laufenden Content-Strecken. Neu ist nicht der Gedanke der Stückkosten selbst. Neu ist, dass KI diesen in Bereiche verlagert, die bislang stärker als individuelle Wissens-, Kreativ- oder Beratungsleistung verstanden wurden. Die eigentliche Frage verändert sich. Es geht immer

noch darum, wie viel eine einzelne Übersetzung, ein einzelnes Asset oder eine einzelne Variante kostet. Es geht aber auch darum, ob ein wiederkehrender Arbeitsschritt so beschrieben, strukturiert und begrenzt werden kann, dass KI ihn verlässlich unterstützt oder teilweise automatisiert. Neue Stückkosten entstehen also dort, wo Arbeit in eine wiederholbare Produktionsarchitektur übersetzt wird.

Diese Architektur besteht aus mehr als einem guten Prompt. Sie braucht saubere Eingangsdaten, klare Vorlagen, definierte Variantenlogiken, Qualitätskriterien, Freigabepunkte, dokumentierte Endpunkte und klare Verantwortlichkeiten. Erst unter diesen Voraussetzungen lässt sich berechnen, was eine Einheit wirklich kostet: in Minuten oder Token, inklusive Prüfung, Steuerung, Datenpflege und Betrieb. Der sichtbarste Hebel dafür ist die Batch-Verarbeitung. Anbieter wie OpenAI und Anthropic gewähren für gebündelte, zeitversetzte Verarbeitung deutliche Preisvorteile. Bei Anthropic liegt der Rabatt aktuell gegenüber synchroner Standardnutzung bei 50 %; auch OpenAI bewirbt seine Batch-Verarbeitung als kostengünstigeres Modell für asynchrone Aufgaben. Aufgaben werden dabei gesammelt, strukturiert und innerhalb eines definierten Zeitfensters abgearbeitet, anstatt dass sie einzeln und sofort bearbeitet werden. Darin liegt kein trivialer Preisvorteil. Es ist vielmehr ein Signal: Wer in größerem Volumen und strukturiert arbeitet, zahlt deutlich weniger. Die Preisarchitektur der Anbieter belohnt organisierte KI-Nutzung. Darin liegt der ökonomische Unterschied zur Einzelabfrage. Eine einzelne KI-Anfrage kann schnell erledigt sein, aber sie schafft noch keine Produktionslogik. Batch-Verarbeitung zwingt zur Vorarbeit: Welche Daten werden verwendet? Welche Felder sind verbindlich? Welche Varianten werden erzeugt? Welche Regeln gelten für Sprache, Tonalität, Format, Länge oder Bildstil? Wie wird geprüft? Was passiert bei Fehlern? Aus dieser Vorarbeit entsteht die eigentliche Skalierungsfähigkeit.

Für Agenturen und Unternehmen ist das gleichermaßen relevant. Unternehmen, die Leistungen intern übernehmen wollen, brauchen genau diese Klarheit und Struktur. Agenturen, die solche Leistungen anbieten, müssen erklären können, ob sie einzelne Ergebnisse liefern oder ein wiederholbares System betreiben. Dieser Unterschied ist entscheidend. Einzelne Ergebnisse werden über Aufwand, während Systeme über Stabilität, Qualität, Durchsatz und Verantwortung bewertet werden.

Standardisierung ist kein Gegenpol von Kreativität, sondern Voraussetzung für skalierbare Entlastung. Nicht jede Leistung kann oder sollte standardisiert werden. Aber dort, wo Arbeit wiederkehrend, regelhaft und qualitätsfähig ist, entscheidet Standardisierung darüber, ob sich KI wirtschaftlich auszahlt oder nur mehr Output erzeugt. Gerade in diesen Feldern entstehen neue Stückkosten: pro Text, pro Variante, pro Analyse, pro Datensatz, pro Prüfpfad, pro Prozesslauf.

Diese neue Stückkostenlogik ist nicht automatisch ein Vorteil für Agenturen. Sie kann Margen verbessern, wenn Produktionsarchitektur, Vertrag und Verantwortlichkeiten zusammenpassen. Sie kann aber auch den Druck erhöhen, wenn Unternehmen dieselbe Logik selbst aufbauen. Deshalb liegt der langfristige Wert in der Fähigkeit, wiederkehrende Arbeit so zu strukturieren, dass sie verlässlich, markenkonform, rechtssicher und wirtschaftlich steuerbar wird. Diese Frage ist größer als Kosteneffizienz.

Das ist auch die Brücke zu agentischen Systemen. Je stärker Prozesse nicht mehr nur einzelne Aufgaben, sondern ganze Arbeitsschritte und Entscheidungsketten umfassen, desto wichtiger wird die Architektur davor. Batch-Verarbeitung ist ein Zwischenschritt: Sie zeigt, dass Skalierung aus Struktur entsteht. Skaleneffekte benötigen Werkzeug und Produktionsarchitektur. Neue Stückkosten entstehen dort, wo KI-Nutzung standardisiert, gebündelt und kontrollierbar wird. Wer diese Architektur beherrscht, kann Leistungen anders kalkulieren. Wer sie nicht beherrscht, bleibt bei schnellerer Einzelarbeit und trägt weiterhin die Unsicherheit des Gesamtprozesses.

Agentische Kosten entstehen nicht nur beim Generieren eines Ergebnisses, sondern **entlang der gesamten Prozesskette.**

Agentic Costs: neue Automatisierung, neue Kostensprünge

Auf Automatisierung folgt Autonomie? Agentische Systeme versprechen genau das: KI soll nicht mehr nur auf einzelne Eingaben reagieren, sondern Aufgaben eigenständig verfolgen, Zwischenschritte planen, Informationen suchen, Werkzeuge nutzen, Ergebnisse prüfen und bei Bedarf nacharbeiten. Aus einem Tool, das unterstützt, wird ein System, das ausführt.

Mit agentischen Systemen verschiebt sich die Arbeitsweise und die Kalkulation. Bei klassischer Automatisierung ist ein Ablauf weitgehend festgelegt. Ein Prozess wird einmal definiert und dann wiederholt. Bei agentischen Systemen ist der Zielrahmen definiert, der Weg dorthin aber dynamischer. Das System kann während der Ausführung weitere Quellen abrufen, zusätzliche Prüfungen anstoßen, Arbeitsschritte wiederholen oder externe Werkzeuge nutzen. Genau darin liegen sowohl sein Nutzen als auch das Kalkulationsrisiko. Bei agentischen Systemen kostet nicht nur das Ergebnis, sondern auch der Weg dorthin: jede Suche, jeder Zugriff, jede Wiederholung, jede Laufzeit, jeder gespeicherte Kontext und jede menschliche Prüfung. Was nach Autonomie aussieht, ist ökonomisch ein laufender Prozess mit variablen Kosten.

Damit verändert sich die Kalkulation. Bei einer einzelnen KI-Anfrage lassen sich die anfallenden Kosten meist grob abschätzen, Batch-Verarbeitung macht sie durch Struktur planbarer. Agentische Systeme machen Kosten hingegen variabler, weil die Ausführung selbst Kosten erzeugt. Ein Agent kann mehr Schritte ausführen als geplant, zusätzliche Quellen abrufen, mehr Kontext speichern, Schleifen wiederholen oder externe Werkzeuge nutzen. Keine dieser Bewegungen ist an sich falsch. Zusammengenommen können sie aber die Kostenbasis deutlich verändern.

Die Preismodelle der Plattformanbieter bilden diese Entwicklung bereits ab. Bepreist wird nicht mehr nur die reine Modellnutzung. Hinzu kommen Suchvorgänge, Dateizugriffe, Laufzeit, Speicher, Codeausführung, Arbeitsumgebungen oder das Speichern und Vorhalten von Kontext. Aus einem Modellpreis wird so der Preis für eine Ausführungsumgebung. Für Agenturen und Unternehmen ist das von Bedeutung: Agentische Kosten entstehen nicht nur beim Generieren eines Ergebnisses, sondern entlang der gesamten Prozesskette: durch Dauer, Häufigkeit, Wiederholungen, Zugriffe und Prüfungen.

Ein beispielhaftes Szenario aus der Praxis verdeutlicht dies: Eine Kommunikationsagentur betreibt für einen Industriekunden einen Themenmonitor, der wöchentlich Fachmedien, Branchenquellen und Wettbewerberseiten auswertet, Meldungen sortiert, Zusammenfassungen erstellt und Empfehlungen für die redaktionelle Planung formuliert. In der ersten Kalkulation werden vor allem die erwartete Menge an Text und Modellnutzung berücksichtigt. Im Betrieb zeigt sich jedoch, dass der Agent deutlich mehr Kosten auslöst: Er führt zusätzliche Suchläufe aus, ruft Quellen mehrfach auf, speichert Kontext, wiederholt Schritte bei unklaren Ergebnissen und wird häufiger gestartet als geplant. Keine dieser Funktionen ist falsch. Zusammen verändern sie aber die Kostenbasis. Was als günstige Automatisierung kalkuliert wurde, wird zu einem laufenden Betrieb mit variablen Kosten.

Notwendig werden Kostenlimits, Abbruchregeln, definierte Laufzeiten, Freigabepunkte, Protokollierung und klare Verantwortlichkeiten. Es muss geklärt werden, wie oft ein Agent laufen darf, welche Quellen er abrufen soll, welche Werkzeuge er nutzen kann, wann er stoppen muss und wann ein Mensch entscheidet. Ohne solche Regeln wird aus Automatisierung schnell ein offener Prozess mit unklaren Kosten.

Für Unternehmen ist das genauso relevant wie für Agenturen. Wer agentische Leistungen einkauft, kauft zusätzlich zur Automatisierung eine Prozessausführung, die laufend Kosten erzeugen kann und überwacht werden muss. Wer solche Leistungen anbietet, muss beschreiben, was der Agent kann und wie sein Betrieb fachlich, technisch und wirtschaftlich begrenzt, geprüft und gesteuert wird. Damit stellt sich nicht nur die Frage nach dem einzelnen Ergebnis und dessen Kosten, sondern auch nach den Kosten des Prozessablaufs. Diese hängen unter anderem davon ab, wie häufig ein Prozess ausgeführt wird, welche Schleifen und Freiheiten erlaubt sind und wer Verantwortung trägt, wenn das System mehr macht als erwartet. Wer heute agentisch arbeitet und die Kosten im Betrieb sieht, versteht, weshalb Plattformanbieter diesen Bereich strategisch bepreisen. Er ist eine künftige Umsatzbasis der Anbieter und zugleich eine neue Kostenbasis ihrer Kunden.

Agentische Systeme können Arbeit entlasten, aber sie machen die Kalkulation nicht einfacher. Sie machen sichtbar, dass KI dort am anspruchs-

vollsten wird, wo sie Werkzeug und Betrieb ist. Diese operative und strategische Dimension wird im Deep Dive Agentic Systems (ab Seite 99 →) vertieft.

KI-Kosten: mehr als eine Toolrechnung

Viele KI-Kalkulationen sind noch aus der Experimentierphase heraus gedacht. Sie rechnen mit Toolpreisen, einzelnen Lizenzen, ersten Produktivitätsgewinnen oder geschätzten Nutzungsvolumen. Was dabei häufig fehlt, ist der Blick auf den laufenden Betrieb: Pflege, Prüfung, Aktualisierung, Freigaben, Datenqualität, rechtliche Absicherung, Überwachung und Verantwortung. Hier entscheidet sich, ob KI-Leistungen wirtschaftlich tragfähig werden.

Eine Automatisierung ist nicht abgeschlossen, wenn sie einmal eingerichtet wurde. Ein KI-gestützter Workflow ist nicht stabil, nur weil er in einem Test funktioniert hat. Ein Assistent bleibt nicht automatisch verlässlich, wenn sich Modelle, Daten, Schnittstellen, Preise oder Nutzungsbedingungen verändern. Professionelle KI-Leistung braucht deshalb laufende Steuerung. KI-Kosten dürfen in Agenturen nicht als verstreute Toolpositionen oder allgemeine Gemeinkosten verschwinden. Nicht jede Lizenz muss einzeln weiterberechnet werden. Aber jede Leistung, die auf KI-Infrastruktur beruht, braucht eine belastbare Kostenlogik, ansonsten werden Aufwände dort unsichtbar, wo sie geschäftskritisch sind: im Projekt, im Kunden-Workflow, im Betrieb gemeinsamer Systeme, in Qualitätssicherung, Datenpflege und Verantwortung.

KI-Kosten dürfen in Agenturen nicht als verstreute Toolpositionen oder allgemeine Gemeinkosten **verschwinden**.

Für Unternehmen intensiviert sich damit die Inhousing-Frage. Wer Leistungen intern übernehmen will, gewinnt möglicherweise mehr Kontrolle, wird aber zugleich mit den Fragen nach Systemen, Nutzungsbedingungen, Datenwegen, Qualität, Betrieb und Verantwortung konfrontiert. Niedrigere Kosten entstehen dadurch nicht automatisch. Aber es ist keine Abkürzung aus der Kosten-, Infrastruktur- und Verantwortungsfrage. Hier liegt eine neue Rolle für Agenturen. Sie können Unternehmen helfen zu unterscheiden, was sinnvoll intern aufgebaut werden sollte, was besser extern betrieben wird und wo gemeinsame Modelle entstehen müssen. Der Wert liegt dann in der Ausführung in Kombination mit der Fähigkeit, KI-gestützte Arbeit wirtschaftlich, rechtlich und organisatorisch steuerbar zu machen. Kalkulation wird zu einer strategischen Aufgabe.

Agenturen müssen erklären können, was eine KI-gestützte Leistung im Betrieb kostet. Unternehmen müssen entscheiden, welche Kosten und Verantwortung sie selbst tragen können und wollen, welche bei einer Agentur verbleiben und welche vielleicht sogar in einer gemeinsamen Infrastruktur entstehen. Erst daraus lässt sich ableiten, ob eine Leistung als Projekt, Pauschale, laufender Betrieb, Lizenzmodell, Beratungsleistung oder gemeinsames System sinnvoll kalkuliert werden kann.

Eine der wichtigsten neuen Fähigkeiten von Agenturen sollte es sein, KI wirtschaftlich steuerbar zu machen. Nur wenn Kosten und Verantwortlichkeiten für einen Betrieb kalkulierbar werden, lässt sich beantworten, was Agenturen künftig eigentlich verkaufen und was Unternehmen tatsächlich einkaufen. Die These lautet deshalb nicht zufällig: KI senkt Kosten nicht, sondern verschiebt diese. Die Frage ist nicht, ob etwas teurer oder billiger wird, sondern, wohin welche Kosten wandern und wer diese am Ende trägt.