

Insight **ON**:AI

Wer steuert Ihre KI?

Was „End-to-End“
wirklich bedeutet

 InsightAI™



Es gibt einen Moment, den viele Technologie-Verantwortliche wiedererkennen werden: Das Projekt-Engagement wird abgeschlossen, das neue System geht live und der Partner wendet sich dem nächsten Projekt zu. Von da an ist die Organisation für etwas verantwortlich, das sie nicht vollständig für den langfristigen Betrieb ausgelegt hat: Governance-Frameworks, die „später“ geregelt werden sollten, Teams, die täglich mit dem System arbeiten, aber nicht an dessen Entwicklung beteiligt waren – und eine Dokumentation, die beschreibt, wie das System gebaut wurde, anstatt wie man es am Laufen hält.

Diese auf die reine Bereitstellung ausgerichteten Partnerschaften sind bei Technologieprojekten nach wie vor üblich. Obwohl sie die Lösung von Problemen am „Tag 2“ besonders schwierig machen, kann die Lücke, die sie hinterlassen, bei manchen Softwarearten überschaubar sein. Bei KI ist dies meist nicht der Fall. Ein KI-System in der Produktion bringt zusätzliche laufende Anforderungen im Vergleich zu herkömmlicher Software mit sich, und diese Anforderungen kündigen sich nicht immer an, wenn etwas schiefzulaufen beginnt.

Viele Organisationen, die KI-Projekte mit Fokus auf die Implementierung angehen, setzen sich erst mit den langfristigen Konsequenzen auseinander, wenn es zu spät ist: Der Partner ist weg, und das System, das sie nun selbst betreiben müssen, erweist sich als deutlich schwieriger zu warten, zu steuern und zuverlässig zu halten als erwartet. Dies ist die Bereitstellungslücke – der Raum zwischen einem eingeführten System und einer KI-Fähigkeit, die zuverlässig und regelkonform läuft. Unternehmen, die diese Lücke erfolgreich schließen, verfolgen oft einen gemeinsamen Ansatz: Sie arbeiten von Anfang an mit Partnern zusammen, die nicht nur die Implementierung, sondern auch die langfristige Betriebsfähigkeit im Blick haben.

Im Jahr 2026 ist das Schließen der Bereitstellungslücke dringender denn je. Ab dem 2. August werden die Pflichten für Hochrisiko-Systeme des EU-KI-Gesetzes (EU AI Act) für in der EU tätige Unternehmen voll durchsetzbar. Die Gesetzgebung verlangt von Organisationen nicht nur, KI verantwortungsvoll zu entwickeln, sondern sie auch im laufenden Betrieb verantwortungsvoll zu betreiben. Dies

bedeutet Konformitätsbewertungen, dokumentierte Prozesse, kontinuierliche Überwachung und klare Rechenschaftspflichten für automatisierte Entscheidungen. „Live gegangen“ ist nicht mehr die Ziellinie, und für Organisationen, deren KI-Partnerschaften um einen Endpunkt bei der Auslieferung herum aufgebaut waren, ist das eine echte Lücke, die es zu schließen gilt.





Die vorläufige Einigung führt auch einen festen Zeitplan für die verzögerte Anwendung von Hochrisiko-Regeln ein: Die neuen Anwendungsdaten wären der 2. Dezember 2027 für eigenständige Hochrisiko-KI-Systeme und der 2. August 2028 für in Produkte eingebettete Hochrisiko-KI-Systeme. Darüber hinaus stellt die vorläufige Einigung die Verpflichtung für Anbieter wieder her, KI-Systeme in der EU-Datenbank für Hochrisiko-Systeme zu registrieren, wenn sie der Ansicht sind, dass ihre Systeme von der Einstufung als Hochrisiko ausgenommen sind. Sie stellt auch den Standard der strikten Notwendigkeit für die Verarbeitung besonderer Kategorien personenbezogener Daten zum Zweck der Erkennung und Korrektur von Verzerrungen (Bias) wieder her.

Die vorläufige Einigung verschiebt die Frist für die Einrichtung von KI-Reallaboren (Regulatory Sandboxes) durch die zuständigen Behörden auf nationaler Ebene auf den 2. August 2027 und verkürzt die Übergangsfrist für Anbieter zur Implementierung von Transparenzlösungen für künstlich generierte Inhalte von 6 Monaten auf 3 Monate, wobei die neue Frist auf den 2. Dezember 2026 festgelegt wurde.

Die Vereinbarung zwischen den Co-Gesetzgebern präzisiert die Zuständigkeiten des KI-Büros bei der Überwachung von KI-Systemen, die auf Basis-KI-Modellen mit allgemeinem Verwendungszweck beruhen – vorausgesetzt, Modell und System stammen vom selben Anbieter. Gleichzeitig werden Ausnahmefälle definiert, in denen nationale Behörden ihre Kompetenzen behalten, insbesondere in den Bereichen Strafverfolgung, Grenzschutz, Justiz und Finanzsektor.

Rechts- und Branchenanalysten betonen, dass die Verzögerung keine „Schlummer-Taste“ für die Compliance ist. Die Kernanforderungen – Data Governance, technische Dokumentation und menschliche Aufsicht – bleiben unverändert. Die primäre Empfehlung des EU-KI-Büros und von Anwaltskanzleien lautet, diese 16-monatige Verlängerung als „Entwicklungsphase“ und nicht als Wartezeit zu betrachten, da die Komplexität der Compliance-Erreichung (insbesondere bei Hochrisiko-Systemen) oft erhebliche betriebliche Überholungen erfordert, deren korrekte Umsetzung über ein Jahr in Anspruch nimmt.

Strategie, Engineering, Operations: Warum KI-Partner alle Bereiche abdecken müssen

Die Bereitstellungslücke tritt so regelmäßig auf, weil die meisten KI-Partnerschaften nur auf eine Phase einer dreiphasigen Reise ausgerichtet sind. Diese unvollständigen Ansätze führen zu drei ausgeprägten Fehlermustern beim langfristigen Erfolg von KI-Projekten.

Das erste Fehlermuster tritt auf, wenn die Beratung bei der Strategie aufhört. Sie erstellen eine schlüssige Roadmap, einen fundierten Business Case und eine wirklich überzeugende Vision davon, was das Produkt leisten könnte. Aber dann wird die Realität des Engineerings zum Problem eines anderen. Die Fragen, die bestimmen, ob ein KI-System in der Produktion tatsächlich funktioniert – wie sich das Modell mit realen Daten statt mit sauberen Testdatensätzen verhält, wie die Inferenzkosten bei Skalierung aussehen und wie es an die bereits im Unternehmen vorhandenen Systeme angebunden wird –, erfordern von Anfang an jemanden im Raum, der diese versteht. Lücken zwischen Strategie und Engineering-Realität können bei jedem Technologieprojekt Probleme verursachen, aber bei KI treten sie oft erst später auf und sind teurer zu beheben.

Das zweite Fehlermuster ist der Bereitstellungspartner, der mit dem Go-Live aufhört. Das System startet, es funktioniert am ersten Tag, und das Projekt wird beendet. Aber ein produktives KI-System ist kein fertiges Produkt. Im Gegensatz zu herkömmlicher Software kann es sich auf eine Weise verschlechtern, die nicht sofort sichtbar ist. Dazu gehören Model Drift (Modell-Drift), Guardrails, die bei veränderten Nutzungsmustern eine

aktive Wartung erfordern, und Ergebnisse, die plausibel erscheinen, aber nicht mehr fundiert sind. Um es gut am Laufen zu halten, sind kontinuierliche Überwachung, aktive Governance und jemand erforderlich, der weiterhin dafür verantwortlich ist, ob die getroffenen Entscheidungen noch die richtigen sind. Wenn jedoch nichts davon im Projektumfang enthalten war, wird dem Unternehmen etwas übergeben, das zwar läuft, aber es wird sich selbst überlassen, wie es diesen Zustand beibehalten soll.

Das dritte Fehlermuster ist im Vorfeld schwerer zu erkennen, in der Praxis jedoch oft ein erhebliches Problem: die reibungslose Integration moderner KI in die Altsysteme (Legacy-Systeme), auf denen das Unternehmen aufbaut. Untersuchungen in The Digital Sovereignty Trilemma ergaben, dass 41 % der IT-Verantwortlichen sagen, dass die Unfähigkeit, geschäftskritische Altanwendungen zu migrieren, sie aktiv ausbremst. Dies ist eine reale, weit verbreitete Herausforderung. Selbst Organisationen, die nicht aktiv durch Altsysteme eingeschränkt sind, verfügen selten über eine völlig neue Infrastruktur (Greenfield). Die meisten werden bereits über Jahrzehnte gewachsene Systeme, Prozesse und Datenarchitekturen

verfügen, die nicht für KI ausgelegt waren. Die Integration von Altsystemen ist auch der Bereich, in dem Bereitstellungspartnerschaften am häufigsten unvollendete Arbeit hinterlassen. Die schwierigsten Verbindungen – zwischen moderner KI-Fähigkeit und den älteren Systemen, die andere Funktionen steuern – werden meist gerade so gut gelöst, dass der Start gelingt, während die tiefergehende Arbeit verschoben wird. Sobald das Projekt beendet ist, fällt diese aufgeschobene Arbeit der Organisation zu.

Eine echte End-to-End-Fähigkeit bedeutet, dass derselbe Partner alle diese Phasen verantwortet. Echte operative Intelligenz setzt sich aus drei Elementen zusammen: einer Strategieentwicklung, die in der Engineering-Realität verankert ist, einem Engineering, das von Tag eins produktionsreif ist, und einer Betriebsstrategie, die von Anfang an integriert ist – um Systeme am Laufen zu halten, compliance-konform zu gestalten und kontinuierlich zu verbessern. Entscheidend ist zudem ein Partner, der dabei hilft, Anwendungsfälle zu identifizieren und zu gestalten, die echten Geschäftswert schaffen – nicht nur solche, die technisch machbar sind.

Aufbau nachhaltiger KI-Fähigkeiten bei einem weltweit führenden Bildungsanbieter

Wie das Schließen der Bereitstellungslücke in der Praxis tatsächlich aussieht, geht weit über das bloße Beherrschen der Technologie hinaus. Wir wurden von einem weltweit führenden Bildungsanbieter kontaktiert, der in fast 200 Ländern tätig ist und digitale Inhalte, Bewertungen und technologiegestützte Lernlösungen auf globaler Ebene anbietet. Die Engineering-Abteilung dieser Organisation kam zu Insight AI mit einer Frage, an die die meisten Unternehmen nicht früh genug denken: Nicht nur, ob KI funktionieren würde, sondern ob ihre eigenen Teams in der Lage wären, sie sicher zu betreiben, richtig zu steuern und selbstständig weiter zu verbessern.

Als InsightAI mit den Engineering-Teams dieser Organisation zusammenarbeitete, offenbarte eine erste Umfrage unter den beteiligten Ingenieuren etwas, das in KI-Adoptions-Schlagzeilen oft verborgen bleibt: Der Großteil des Teams nutzte KI-Tools bereits regelmäßig – die Zahlen sahen also gut aus. Doch dies ist ein Muster, das sich in vielen Unternehmen zeigt: Regelmäßige Nutzung von KI-Tools bedeutet nicht automatisch effektive

oder sichere Nutzung. Eine detaillierte Analyse der tatsächlichen Praktiken zeigte das eigentliche Problem: Die Bewertung von KI-Ausgaben erfolgte unstrukturiert und inkonsistent, während anspruchsvollere Techniken – die typischerweise die größten Gewinne bringen – deutlich zu wenig eingesetzt wurden.

Noch bedeutender war das, was die Menschen zurückhielt. Die Angst um Sicherheit und Compliance war die am häufigsten genannte Barriere, die von Ingenieuren über alle Rollen und Funktionen hinweg angeführt wurde. Mitarbeiter, die sich unsicher waren, wählten bestimmte Tools entweder vollständig oder trafen eigene Ermessensentscheidungen ohne ein konsistentes Framework, das sie leitete. Das Problem war nicht Unachtsamkeit, sondern das Fehlen von Standards, Tool-Richtlinien und dem nötigen Vertrauen, um sicher zu arbeiten. Der Einsatz von noch mehr Technologie würde das nicht lösen.



Was sich geändert hat

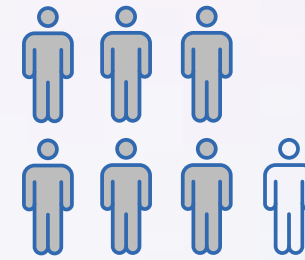
Das Programm zielte darauf ab, das aufzubauen, was den Teams gefehlt hatte: die Infrastruktur für eine sichere und effektive KI-Nutzung. Dies geschah durch eine Kombination aus Workshops, die auf realer Engineering- und QA-Arbeit basierten, eingebettetem Coaching in Live-Codebasen und Testsuiten sowie einer Reihe von Governance-Artefakten – Nutzungsleitfäden, Entscheidungsrahmen, Prompt-Bibliotheken, Best-Practice-Playbooks –, die explizit darauf ausgelegt waren, das Programm zu überdauern. Anstatt diese Standards nach Beendigung des Projekts wieder in den Hintergrund treten zu lassen, wurde ein internes Champion-Netzwerk aufgebaut, bei dem die erfahrensten Praktiker die sichtbare Verantwortung für die neu entstehenden Standards und eine Lehrrolle in der Coaching-Phase übernahmen.

Der Governance-Wert ist derjenige, auf den es im Kontext der Frist am 2. August am meisten ankommt. Am Ende des Programms nannten leitende Mitarbeiter spezifische Kontrollen: bestätigte Enterprise-Lizenzkonfigurationen, anonymisierte Snippets für sensiblen Code und Pull-Request-Offenlegung (PR) als Standardpraxis. Ein Team fing ein schreibberechtigtes Tool ab, das Ticket-Inhalte überschrieb, bevor es zu einem Live-Vorfall wurde – die Art von praktischem Risikobewusstsein, die nicht aus einem Richtliniendokument stammt und deren Aufbau Monate dauert.

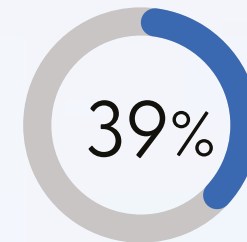
„Lies jede Zeile. KI generiert plausibel aussehenden Code, der etwas völlig Falsches testen kann. Wenn du den Test mit einer schlechten Eingabe nicht fehlschlagen lassen kannst, testet er gar nichts.“

QA Engineer

Die Ergebnisse gingen weit über das hinaus, was zu Beginn angestrebt worden war:



6 von 7 Entwicklern sagen jetzt, dass KI-Tools ihnen bei der tatsächlichen Arbeit, die sie abliefern, echte Zeit einsparen



39% Reduzierung der Durchlaufzeit bis zum ersten Code-Commit, verglichen mit einem Ziel von 15–20 %



2–8 Stunden Zeitersparnis pro erstelltem automatisierten Test



10–30 Minuten Zeitersparnis pro Fehlerbericht (Bug Report)



Von der Adoption zur Eigenständigkeit (Agency)

Der stärkste Beweis dafür, dass diese Fähigkeit tatsächlich übertragen wurde, zeigte sich nach dem Ende des Programms. Das QA-Team, die Gruppe, die sich zu Beginn selbst am konservativsten eingeschätzt hatte, entwickelte in der Folge zwei Agenten, die gezielt Probleme in ihrem eigenen Bereitstellungsprozess angingen: Einer reduzierte die manuelle Arbeit zur Einhaltung der Barrierefreiheit von bis zu 38 Stunden pro Seite auf rund 30 Minuten, ein anderer verkürzte die Zeit für die Design-Überprüfung pro Komponente von bis zu drei Stunden auf unter zehn Minuten. Keiner von beiden wurde von InsightAI gebaut. Sie wurden von den eigenen Ingenieuren des Teams gebaut, mit Fähigkeiten, die nun dieser Organisation gehören.



Möchten Sie diese Art von interner KI-Kompetenz in Ihren eigenen Teams aufbauen?

Unsere Spezialisten können Sie durch den Ansatz dieses Programms führen – wie Kompetenztransfer, Governance und reale Anwendung so strukturiert wurden, dass sie über die Projektlaufzeit hinaus Bestand haben.

[Sprechen Sie mit einem Spezialisten für KI-Kompetenz]

Engineering mit Blick auf das Ziel

Dieses Programm zeigt, was es bedeutet, nachhaltige Fähigkeiten auf organisatorischer Ebene aufzubauen – indem man die Menschen, die ein KI-System betreiben, in die Lage versetzt, es sicher zu nutzen und gut zu steuern. Dieselbe Frage stellt sich auf der Engineering-Ebene: Ob das System selbst so gebaut wurde, dass es im Laufe der Zeit betrieben, angepasst und gesteuert werden kann, oder ob es einfach nur für den Start gebaut wurde.

Wir haben diese Prinzipien selbst in die Praxis umgesetzt, als wir AURA entwickelten – eine KI-gestützte Wissensplattform zur Lösung eines echten internen Problems: Die Umwandlung abgeschlossener Projektarbeiten in kundenorientierte Fallstudien war zeitaufwändig, manuell und führte regelmäßig dazu, dass die Inhalte nicht rechtzeitig vorlagen, wenn sich eine Vertriebschance bot. Mit AURA haben unsere Vertriebsteams unmittelbaren Zugriff auf relevante Belege während ihrer Deal-Arbeit und können damit Case Studies in mehreren Sprachen erstellen, die unsere Expertise demonstrieren – mit direktem kommerziellem Gewinn.

Der Aufbau gab uns jedoch auch die Möglichkeit, dieselben Engineering-Prinzipien auf unsere eigenen Systeme anzuwenden, die wir auch bei einem Kundenprojekt anwenden würden. Das Prinzip hinter diesen Entscheidungen ist einfach: Die Entscheidungen, die bestimmen, ob man einem System im Laufe der Zeit vertrauen kann, sind in einer Demo selten sichtbar. Sie zeigen sich Monate später, wenn die vom System verarbeiteten Daten sensibler werden, wenn die Nutzung skaliert, wenn sich die regulatorische Landschaft verschärft.

Anstatt Governance und Compliance als nachträgliche Ergänzungen zu behandeln, haben wir sie von Anfang an integriert. Daten werden standardmäßig anonymisiert, der Zugriff ist rollenbasiert geregelt. Infrastrukturänderungen sind auditierbar und reproduzierbar – nicht manuell angewendet und schwer nachvollziehbar. Und das operative Monitoring ist von Anbeginn integriert, nicht nachträglich hinzugefügt.

Keine dieser Entscheidungen ändert etwas daran, was das System am ersten Tag tut, aber sie bestimmen, ob



AURA verantwortungsvoll im großen Maßstab betrieben, bei veränderten Anforderungen angepasst und unter zunehmend anspruchsvollen Bedingungen gesteuert werden kann – einschließlich der Verpflichtungen, die das EU-KI-Gesetz nun fortlaufend statt einmalig vorschreibt. AURA läuft nach wie vor, hat sich weiterentwickelt und wird von denselben Fundamenten gesteuert, die wir zu Beginn gelegt haben. Die Plattform hat sich mittlerweile so weiterentwickelt, dass sie breitere themen- oder branchenfokussierte Narrative aus Gruppen relevanter Fallstudien erstellen kann – etwas, das auf einem System, das nicht auf Wachstum ausgelegt war, unmöglich zu bauen gewesen wäre.

Was der August 2026 tatsächlich erfordert

Die Hochrisiko-Pflichten des EU-KI-Gesetzes können nicht durch eine einmalige Abnahme erfüllt werden. Sie erfordern fortlaufende Konformität: kontinuierliche Dokumentation, Überwachung, menschliche Aufsicht über automatisierte Entscheidungen und klare Rechenschaftsstrukturen, die nachgewiesen statt nur beschrieben werden können. Für Organisationen, deren KI-Partnerschaften mit der Auslieferung enden, wird die Erfüllung dieses Standards eine betriebliche Herausforderung sein, auf die sie nicht vorbereitet sind.

Was das KI-Kompetenzprogramm des Bildungsanbieters veranschaulicht, ist, wie regulatorische Bereitschaft in der Praxis aussieht: die 90 % der leitenden Mitarbeiter, die spezifische Kontrollen benennen und im Kontext anwenden können; das Team, das ein Governance-Risiko abfing, bevor es zu einem Vorfall in der Produktion wurde; die Standards, Governance-Materialien und das Champion-Netzwerk, die die KI-Nutzung der Organisation weiterhin steuern, während sich die Technologie weiterentwickelt.

Diese Art von operativem Fundament wird zunehmend auch zu einem kommerziellen Vorteil. Die Untersuchung **The Digital Sovereignty Trilemma** ergab, dass 55 % der Unternehmen die regulatorische Komplexität – DSGVO, DORA, das KI-Gesetz – als eine ihrer größten strategischen Herausforderungen nennen. Wo die meisten auf Schwierigkeiten stoßen, gewinnen diejenigen, die gut damit umgehen, einen Vorteil: 60 % der Kunden fordern mittlerweile im Rahmen der Beschaffung einen Compliance-Nachweis, und 43 % der Organisationen haben Souveränitätsnachweise genutzt, um Aufträge zu gewinnen oder zu halten. Die Organisationen, die eine kontinuierliche Compliance nachweisen können und sich nicht nur auf eine Abnahme bei der Auslieferung berufen, sind diejenigen, die im Zuge wachsender Verpflichtungen im Vorteil sein werden.



Die volle Distanz

Das KI-Kompetenzprogramm unseres Bildungsanbieters und unser AURA-Projekt zeigen beide, wie nachhaltiger Erfolg in der Praxis aussieht – allerdings aus unterschiedlichen Perspektiven. Das KI-Kompetenzprogramm baute interne Fähigkeiten auf: Standards, Governance und das Vertrauen, sich kontinuierlich zu verbessern. Dies ermöglichte der Organisation, diese Fähigkeiten nach unserem Ausscheiden eigenständig weiterzutragen. AURA hingegen wurde von Anfang an auf Langlebigkeit ausgelegt – die Entscheidungen, die langfristige Betriebsfähigkeit bestimmen, wurden während der Entwicklung getroffen, nicht auf später verschoben. Beides lief auf dieselbe Frage hinaus: Wer denkt vor Beginn der Bereitstellung über die Zeit nach der Bereitstellung nach?

Für die meisten Organisationen wird diese Frage lange vor dem Live-Gang des Systems beantwortet. Sie hängt davon ab, ob der Partner, der es gebaut hat, von Anfang an an den Betrieb gedacht hat: Ob die Strategie die Realität des Engineerings berücksichtigte, ob das Engineering für das ausgelegt war, was betrieben werden musste, und ob die fortlaufende Verantwortung Teil des Projekts war, anstatt ein Gespräch für später zu sein.

Diese Kontinuität der Verantwortlichkeit – über Strategie, Engineering und die laufende Arbeit, ein System am Laufen zu halten – ist das, was End-to-End tatsächlich bedeutet.

Möchten Sie sehen, wie sich diese Governance- und Engineering-Prinzipien in einem Live-System widerspiegeln?

Wir können Sie durch das Design von AURA für langfristige Betriebsfähigkeit führen, einschließlich Compliance-by-Design, Monitoring und skalierbarer Architekturwahlen.

[Fordern Sie eine AURA-Präsentation an]

Lesen [Sie Immer noch am Pilotieren? So bringen Sie KI schneller in die Produktion](#) dieser Herausforderung – wie man KI vom Pilotprojekt in die Produktion überführt und was es braucht, um dorthin zu gelangen.

ch.insight.com/ai

