KI-Allianz

KI-PROJEKTBÜRO

Monatlicher KI-Hintergrund-Newsletter

KI-Allianz Rheinland-Pfalz

KI-Allianz Rheinland-Pfalz - Monatlicher KI-Hintergrund-Newsletter

September 2026

KI-Souveränität für KMU: Wo sollte Ihre KI laufen?

Warum KI-Souveränität wichtig ist

KI-Souveränität klingt zunächst nach einem strategischen Thema für Regierungen und Großunternehmen, wird aber auch für KMU zu einer sehr praktischen Frage. Wo Ihre KI läuft, beeinflusst Kosten, Verfügbarkeit, Leistung, Datenschutz, Geschäftsgeheimnisse, Anbieterabhängigkeit und die Frage, wie leicht Sie später die Technologie wechseln können.

Europa misst technologischer Souveränität bei KI, Cloud, Halbleitern und Open Source zunehmend Bedeutung bei. Für ein KMU bedeutet Souveränität jedoch nicht, jede nicht-europäische Komponente zu vermeiden. Entscheidend ist zu verstehen, welche Teile des KI-Stacks Sie kontrollieren - und den Grad an Kontrolle passend zur Sensibilität und Bedeutung des Anwendungsfalls zu wählen.

Souveränität ist mehr als der Serverstandort

Daten & Geschäftsgeheimnisse
Wohin gehen Prompts, Dokumente, Logs, Embeddings und Modellausgaben - und wer kann technisch darauf zugreifen?

Rechtsraum
Welches Unternehmen ist Ihr Vertragspartner, welche Rechtsordnungen können anwendbar sein und sind internationale Datenübermittlungen beteiligt?

Betriebliche Kontrolle
Können Sie weiterarbeiten, wenn Internet, API, Anbieter oder ein bestimmtes Modell ausfällt?

Technologie & Lock-in
Können Sie Modelle, Daten, APIs und Workflows anderswohin verschieben, ohne alles neu aufzubauen?

Wirtschaftlichkeit & Kompetenzen
Bevorzugen Sie Investitionen in eigene Hardware und Eigenbetrieb - oder variable Cloud-Kosten und Managed Services?

Auch lokale KI hängt von globaler Hardware, Software, Modellen und Lieferketten ab. Ziel ist nicht absolute Unabhängigkeit, sondern bewusste Kontrolle über die Abhängigkeiten, auf die es ankommt.

Vom Laptop bis zur globalen Cloud

Es gibt nicht das eine "souveräne" Bereitstellungsmodell. Das folgende Spektrum zeigt sieben typische Möglichkeiten, wie ein KMU KI betreiben kann - von vollständig lokaler Inferenz bis zu global betriebenen KI-Diensten - und die wichtigsten Zielkonflikte auf jeder Stufe.

Sieben Bereitstellungsstufen - wichtigste Vorteile und Zielkonflikte

Die folgenden Stufen sind keine Compliance-Rangliste. Sie zeigen, wie sich operative Kontrolle, Elastizität, rechtliche Exposition, Kostenstruktur und technische Abhängigkeit typischerweise verändern, je mehr Teile des Stacks ein externer Anbieter betreibt.

Mehr direkte Kontrolle, offline, feste Kapazität — **Mehr Elastizität, Managed Services, neueste Modelle**



Souveränität ist ein Spektrum - kein Ja/Nein-Label

Stufe	Typischer Aufbau / Beispiele	Wesentliche Vorteile	Wesentliche Zielkonflikte
1 Lokale KI auf Gerät oder Workstation	MacBook mit Apple Silicon, GPU-Workstation, NVIDIA DGX Spark, spezielle KI-Box	Maximale direkte Kontrolle über Daten und Geschäftsgeheimnisse; offline nutzbar; sehr geringe Latenz; nach Hardwarekauf keine Gebühren pro Token oder Anfrage.	Durch lokalen Speicher und Rechenleistung begrenzt; größere Modelle ggf. langsam oder nicht möglich; Hardware altert; Patches, Modellverwaltung, Backup und Endpoint-Sicherheit bleiben Ihre Aufgabe.
2 Eigener GPU-Server On-Premises	Serverraum im Unternehmen mit einer oder mehreren GPUs; zentraler lokaler Inferenz- oder RAG-Dienst	Sensible Daten können im Unternehmensnetz bleiben; mehr Kapazität als auf einer Workstation; zentrale Governance für mehrere Nutzende und Anwendungen.	Hohe Anfangsinvestition; Strom, Kühlung, Redundanz, Wartung und Spezialwissen erforderlich; Kapazität kann ungenutzt bleiben oder bei Lastspitzen zu klein sein.
3 Eigene Hardware im Colocation-Rechenzentrum	Unternehmenseigener GPU-Server oder Rack in einem EU-Rechenzentrum	Hardware und Software bleiben unter eigener Kontrolle; professionelle Infrastruktur liefert Strom, Kühlung, physische Sicherheit und Konnektivität; größere Systeme leichter betreibbar als im Büro.	Hardware-Investition und Lebenszyklusrisiko bleiben; Fernadministration und Rechenzentrumsgebühren erhöhen die Komplexität; Skalierung langsamer als bei gemieteter Cloud-Kapazität.
4 Cloud-/KI-Anbieter mit Hauptsitz in der EU	Beispiele: IONOS, OVHcloud, Scaleway, STACKIT	GPU-Kapazität nach Bedarf ohne eigene Hardware; europäischer Vertragspartner und Datenresidenzoptionen können Souveränitäts- und Datenschutzfragen vereinfachen; einfachere Skalierung und Managed Services.	Compliance hängt weiterhin von Konfiguration und Anwendungsfall ab; Portfolio ggf. kleiner als bei globalen Hyperscalern; bei kontinuierlicher GPU-Nutzung können Cloud-Kosten schnell steigen.
5 Souveräne europäische Varianten globaler Cloud-Technologie	Beispiele: AWS European Sovereign Cloud; partnerbetriebene Google Cloud Souveränitätsangebote wie S3NS oder T-Systems	Breites Technologie-Ökosystem kombiniert mit stärkeren EU-bezogenen Betriebs-, Residenz- und Governance-Kontrollen; sinnvoll, wenn vertraute Hyperscaler-Technologie benötigt wird, aber höhere Souveränitätsanforderungen bestehen.	Souveränität variiert je nach Dienst und Vertragsstruktur; nicht jeder Dienst ist verfügbar; zugrunde liegende Technologie und Konzern-/Lieferkette können weiterhin Abhängigkeiten außerhalb Europas schaffen.
6 Standard-EU-Region / europäische Gesellschaft eines US-Hyperscalers	AWS, Microsoft Azure oder Google Cloud für EU-Regionen bzw. EU-Datengrenzen konfiguriert	Sehr breites Angebot, ausgereifte Werkzeuge, globaler Support, elastische Skalierung und schneller Zugriff auf neue KI-Funktionen; Kontrollen zur EU-Datenresidenz werden immer umfangreicher.	EU-Datenstandort ist nicht dasselbe wie vollständige rechtliche oder operative Souveränität; Support, Telemetrie, Unterauftragnehmer oder Ausnahmeübermittlungen können weiter zu prüfen sein; Lock-in kann erheblich sein.
7 Globaler bzw. US-KI-Cloud-Dienst / API	Direkte SaaS- oder Frontier-Modell-APIs, global oder aus den USA betrieben	Schnellster Zugang zu führenden Modellen und Funktionen; nahezu kein Infrastrukturaufwand; sehr gut für Experimente und stark schwankende Lasten.	Geringste Infrastrukturkontrolle; Verträge, Speicherfristen, Nutzung zum Modelltraining, internationale Übermittlungen, Verfügbarkeit, Preisänderungen und Anbieterabhängigkeit müssen genau geprüft werden - besonders bei sensiblen Daten.

Wichtig: "europäisch", "in der EU gehostet", "DSGVO-konform" und "souverän" sind keine Synonyme. Prüfen Sie den konkreten Dienst, die Rechtseinheit, das Betriebsmodell, Supportzugriffe, Unterauftragnehmer und Datenflüsse - nicht nur den Standort des Rechenzentrums.

Was ändert sich entlang des Spektrums?

Vom Level 1 in Richtung 7 steigen in der Regel Elastizität, Managed Services und Zugang zu neuesten Modellen. In der Gegenrichtung nehmen direkte Kontrolle, Offline-Fähigkeit und Unabhängigkeit von externen Diensten zu. Keine Richtung ist automatisch besser: Die passende Architektur hängt davon ab, was die KI tut und welche Folgen ein Datenabfluss, geänderte Anbieterbedingungen oder eine Nichtverfügbarkeit des Dienstes hätten.

Eine wichtige Nuance: Auch ein vollständig lokales Setup ist nicht vollständig "europäisch" oder unabhängig. GPUs, CPUs, Firmware, Betriebssysteme, Modellgewichte und Softwarebibliotheken sind Teil globaler Lieferketten. Souveränität ist deshalb ein Kontinuum - gute Architektur macht Abhängigkeiten sichtbar, statt sie auszublenden.

Hybrid ist oft die praktische Lösung

Für viele KMU ist eine hybride Architektur die praktikabelste Antwort. Sensible Dokumente, Embeddings, Geschäftsgeheimnisse oder eine Unternehmenswissensbasis können lokal oder auf kontrollierter europäischer Infrastruktur bleiben, während große Cloud-Modelle nur für Aufgaben eingesetzt werden, die die sensibelsten Informationen nicht benötigen.

Schon eine einfache Routing-Regel kann helfen: öffentliche oder wenig sensible Aufgaben nutzen einen flexiblen Cloud-Dienst; vertrauliche Aufgaben gehen in eine EU-kontrollierte Umgebung; hochsensible Workloads bleiben On-Premises. Standard-APIs und portable Modellformate erleichtern später einen Anbieterwechsel.

Sieben Fragen vor der Anbieterwahl

1. Wo werden Prompts, hochgeladene Dateien, Logs, Embeddings und Ausgaben gespeichert und verarbeitet?
2. Welche Rechtseinheit schließt den Vertrag und welche Muttergesellschaft oder Nicht-EU-Rechtsordnung kann weiterhin relevant sein?
3. Können Beschäftigte des Anbieters oder Unterauftragnehmer auf die Daten zugreifen - und aus welchen Ländern?
4. Werden Kundendaten zum Training oder zur Verbesserung gemeinsamer Modelle genutzt, und lässt sich dies vertraglich ausschließen?
5. Kann Modell oder Workload später anderswo betrieben werden, oder ist er an proprietäre APIs, Datenbanken oder Orchestrierung gebunden?
6. Wie hoch sind die realen Kosten bei Dauerlast: Hardware, Strom, Administration, GPU-Stunden, Tokens, Speicher und Datentransfer?
7. Was ist der Fallback, wenn Modell, API, Internetverbindung oder Anbieter nicht verfügbar ist?

Regulatorischer Realitätscheck

Lokales Hosting macht eine KI-Nutzung nicht automatisch DSGVO- oder AI-Act-konform; umgekehrt macht die Nutzung eines US-Anbieters sie nicht automatisch unzulässig. Entscheidend bleiben Daten, Zweck, Ihre Rolle, Risikokategorie, Vertrag und ein ggf. erforderlicher Mechanismus für internationale Übermittlungen. Für Übermittlungen an zertifizierte US-Organisationen ist das EU-US Data Privacy Framework ein möglicher Mechanismus; je nach Fall kommen auch andere Garantien in Betracht.

Auch im Blick behalten: der EU Data Act

Vendor-Lock-in ist auch regulatorisch relevant. Der EU Data Act enthält Regeln für Cloud-Wechsel und Interoperabilität. Ab 12. Januar 2027 dürfen Anbieter von Datenverarbeitungsdiensten für den Wechselprozess keine Wechselentgelte mehr erheben; die EU-Kommission zählt hierzu auch relevante Data-Egress-Gebühren. Das macht nicht jeden Workload portabel, aber Exit-Planung wird damit zu einer zunehmend sinnvollen Beschaffungsanforderung.

Ausblick

Eine sinnvolle Souveränitätsstrategie beginnt nicht mit der Frage "Cloud oder keine Cloud?". Sie beginnt mit einer Übersicht über Datensensibilität, erforderliche Modellleistung, Verfügbarkeit, Kosten und akzeptable Abhängigkeiten. Sind diese geklärt, können KMU bewusst lokal, On-Premises, in einer europäischen Cloud, in einer globalen Cloud - oder in einer Kombination daraus - arbeiten.

Inhalte zusammengestellt vom KI-Projektbüro der KI-Allianz Rheinland-Pfalz,
unterstützt durch den Lehrstuhl Maschinelles Lernen an der RPTU in Kaiserslautern.
Steffen Reithermann <steffen.reithermann@cs.rptu.de> - ml.cs.rptu.de
Newsletter-Archiv: https://ml.cs.rptu.de/projects/KI-Projektbuero/newsletter/
Virtuelle Plattform der KI-Allianz Rheinland-Pfalz: https://www.ki-allianz-rlp.de/
Website des KI-Projektbüros: https://ml.cs.rptu.de/projects/KI-Projektbuero/

RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau