Digital **Insight**

Mai 2026

Souveränität als Türöffner – KI-Architektur zwischen Anspruch und Realität

Souveränität als architektonische Eigenschaft

Digitale Souveränität ist im öffentlichen Sektor weniger ein strategisches Leitmotiv als eine operative Voraussetzung. Spätestens mit dem produktiven Einsatz generativer KI entscheidet sie darüber, ob Projekte überhaupt genehmigt, umgesetzt und langfristig betrieben werden können. Zwischen politischen Zielbildern, steigenden Fallzahlen, Fachkräftemangel und strengen rechtlichen Rahmenbedingungen entsteht ein Spannungsfeld, das sich nicht durch Grundsatzentscheidungen auflösen lässt. Es erfordert konkrete architektonische Antworten.

Gerade im KI-Kontext zeigt sich, dass viele Vorhaben nicht an fehlender fachlicher Vision scheitern, sondern an ungeklärten Fragen zu Betrieb, Kontrolle und Verantwortung. Digitale Souveränität wird dabei häufig als abstraktes Schutzversprechen gedacht, ohne präzise zu definieren, wo sie technisch konkret verankert ist. Der folgende Beitrag betrachtet Souveränität daher nicht normativ, sondern aus der Perspektive realer Systemarchitektur.

Kontrollierte KI: Architekturprinzipien für souveräne Assistenzsysteme

Ausgangspunkt der folgenden Betrachtung ist ein konkretes Projekt zur Entwicklung eines KI-basierten Assistenzsystems für das deutsche Sozialrecht. Ziel des Projektes war es, Sachbearbeitende und juristische Fachkräfte bei der täglichen Arbeit zu unterstützen, ohne Entscheidungen zu automatisieren oder Verantwortung zu verlagern. Der Mehrwert liegt in der strukturierten Aufbereitung komplexer Rechtsmaterie: relevante Normen, Auslegungshilfen, Verwaltungsvorschriften und aktuelle Rechtsprechung sollen fallbezogen auffindbar, vergleichbar und nachvollziehbar bereitgestellt werden.

Technologisch basiert die Lösung auf einer Retrieval-Augmented-Generation-Architektur. Zentrales Designprinzip ist die konsequente Trennung zwischen Sprachmodell, fachlichem Wissensbestand und Steuerungslogik. Große Sprachmodelle kommen ausschließlich für die sprachliche Generierung und Strukturierung zum Einsatz. Fachliche Inhalte werden nicht in Modellen angelernt, sondern

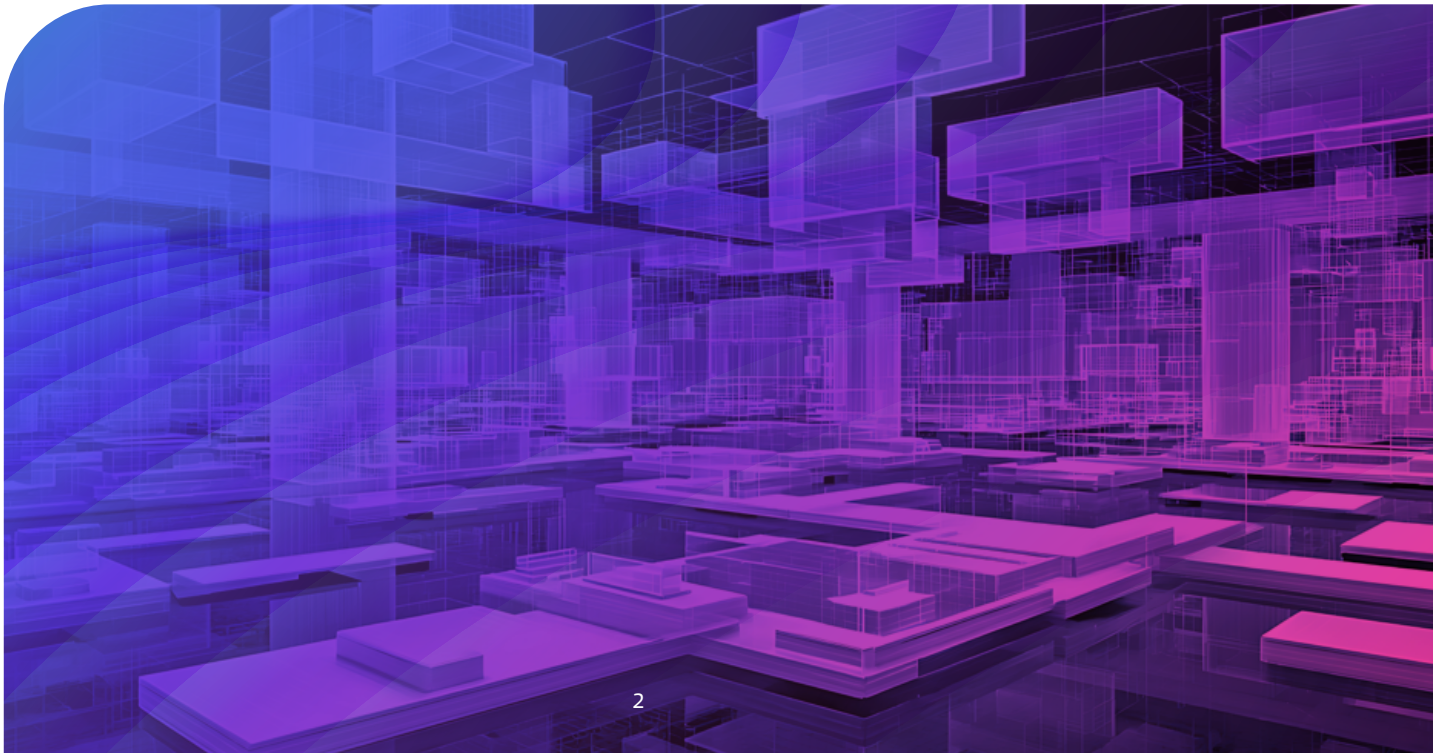
kontrolliert aus Vektor- und Dokumentenspeichern eingebunden. Damit bleibt nicht nur die fachliche Hoheit erhalten, sondern auch die Möglichkeit, Quellen, Versionen und Aktualität jederzeit nachzuvollziehen.

Eine zentrale Designentscheidung ist die strikte Begrenzung der Modellverantwortung. Das System ist bewusst nicht als autonomes Expertensystem ausgelegt, sondern als Assistenz mit klar definierten Schnittstellen zur menschlichen Entscheidung. Antworten werden stets mit referenzierten Quellen versehen, Unsicherheiten explizit gemacht und eine finale Bewertung bewusst dem Menschen überlassen. Diese Architektur ist wesentlich, um Akzeptanz bei Fachanwendenden herzustellen.

Bereits in frühen Projektphasen zeigte sich jedoch, dass nicht das fachliche Konzept, sondern der Infrastrukturdiskurs die größte Hürde darstellte. In vielen öffentlichen Organisationen gilt der On-Premises-Betrieb weiterhin als impliziter Standard für souveräne IT. In der praktischen Umsetzung erwies sich dieses Paradigma jedoch als kaum tragfähig. Insbesondere KI-Workloads mit GPU-basierter Inferenz stellen Anforderungen an Skalierbarkeit, Redundanz, Monitoring und Wartung, die klassische Rechenzentrumsansätze nur mit erheblichem Aufwand erfüllen können.

Dem gegenüber stehen klassische Public-Cloud-Modelle, die technische Skalierbarkeit und Automatisierung bieten, jedoch an anderen Stellen scheitern. Der öffentliche Sektor bewertet Cloud-Risiken nicht primär entlang einzelner Datenschutzparagrafen, sondern anhand struktureller Kontrollmöglichkeiten. Globale Control-Planes, nicht-europäische Betriebsteams, potenzielle Fremdzugriffe und schwer trennbare Metadatenflüsse standen im Widerspruch zu den politischen Zielbildern digitaler Souveränität.

Im Projekt wurde daher deutlich, dass digitale Souveränität keine juristische Kategorie ist, sondern eine technische Eigenschaft von Architekturen. Entscheidend ist nicht der Vertragsstandort eines Anbieters, sondern die Frage, wer operativ Zugriff hat, wie dieser technisch abgesichert ist und ob sämtliche Betriebs- und Änderungsprozesse revisionssicher nachvollzogen werden können. Moderne KI-Architekturen benötigen dafür integrierte Plattformdienste, etwa für Identitätsmanagement, Security-Automatisierung, Logging und Compliance-Kontrollen.



Umsetzung in der Praxis: Warum weder europäischer Anbieter noch Open Source

In der praktischen Umsetzung stellte sich früh die Frage, ob digitale Souveränität nicht konsequenter durch den Einsatz europäischer Cloud-Anbieter oder durch einen vollständig Open-Source-basierten KI-Stack erreicht werden könnte. Beide Optionen wurden im Projekt systematisch geprüft, erwiesen sich jedoch aus unterschiedlichen Gründen als nicht tragfähig.

Europäische Cloud-Anbieter bieten zwar konforme Basissysteme, verfügen jedoch nicht über die notwendige funktionale Tiefe für generative KI. Insbesondere skalierbare GPU-Kapazitäten, produktionsreife Vektor-Services sowie integrierte Security-Automation fehlten. Die Konsequenz wären fragmentierte Architekturen mit hohem Integrations- und Betriebsaufwand gewesen.

Auch ein vollständig Open-Source-basierter Ansatz wurde intensiv evaluiert. Zwar bietet Open Source maximale Transparenz auf Code-Ebene, verlagert aber Betrieb, Skalierung, Sicherheit und Compliance vollständig auf den Betreiber. Für eine produktive Nutzung im öffentlichen Sektor hätte dies dauerhaft erhebliche personelle und organisatorische Ressourcen gebunden und vermeidbare Risiken erzeugt.

Vor diesem Hintergrund erwies sich die AWS European Sovereign Cloud als tragfähiger architektonischer Kompromiss. Ausschlaggebend war nicht die Herkunft des Anbieters, sondern die konsequente Trennung von Daten-, Betriebs- und Zugriffsebenen. Betrieb durch ausschließlich EU-basiertes Personal, isolierte Control-Planes und hardwaregestützte Zugriffsbeschränkungen ermöglichen ein Betriebsmodell, das technologische Leistungsfähigkeit mit regulatorischer Kontrolle verbindet.

Souveräne KI als Ergebnis architektureller Entscheidungen

Das Projekt zeigt exemplarisch, dass digitale Souveränität im Kontext Künstlicher Intelligenz nicht durch Abschottung oder technologische Selbstbeschränkung entsteht, sondern durch bewusste architektonische Entscheidungen. Der reflexhafte Verzicht auf cloudbasierte Plattformen führt dabei nicht zu mehr Kontrolle, sondern verlagert Risiken in den operativen Betrieb und begrenzt Innovationsfähigkeit dort, wo sie im öffentlichen Sektor dringend benötigt wird.

Entscheidend ist, Souveränität nicht als juristische Kategorie, sondern als überprüfbare Systemeigenschaft zu begreifen. Wer KI-Systeme verantwortungsvoll einsetzen will, muss nachvollziehbar beantworten können, wo Daten liegen, wie Modelle betrieben werden, welche Zugriffsmöglichkeiten bestehen und wie sich diese Strukturen auditieren lassen. Erst wenn diese Fragen technisch sauber gelöst sind, wird Souveränität belastbar – unabhängig davon, ob eine Lösung formal als Cloud- oder On-Premises-System betrieben wird.

Souveräne Cloud-Modelle eröffnen dem öffentlichen Sektor neue Handlungsspielräume. Sie ermöglichen es, leistungsfähige KI-Architekturen zu nutzen, ohne zentrale Kontroll- und Governance-Anforderungen aufzugeben. Voraussetzung ist jedoch, dass diese Modelle nicht als Abkürzung verstanden werden, sondern als integraler Bestandteil einer übergeordneten Architektur- und Governance-Strategie. Verantwortung, Betrieb und Weiterentwicklung müssen klar geregelt und dauerhaft überprüfbar bleiben.

Für Entscheidende im öffentlichen Sektor ergibt sich daraus eine klare Schlussfolgerung: Die zentrale Frage lautet nicht mehr, ob Cloud oder ob KI, sondern wie diese Technologien so gestaltet werden, dass staatliche Handlungsfähigkeit, technologische Leistungsfähigkeit und regulatorische Kontrolle zusammengeführt werden. Digitale Souveränität wird damit vom politischen Schlagwort zur operativen Voraussetzung für zukunftsfähige KI-Anwendungen – und letztlich für die Modernisierungsfähigkeit staatlicher IT insgesamt.

AUTOREN

Richard Pielczyk,
Abteilungsleiter Public Sector, msg
Richard.Pielczyk@msg.group

Pascal Hinrichs,
Senior Business Consultant
Public Sector, msg
Pascal.Hinrichs@msg.group

IMPRESSUM

Herausgeber
msg systems ag
Robert-Bürkle-Straße 1
85737 Ismaning/München
Deutschland

Redaktionsleitung:
Lennard Munschke
msg systems ag
Rummelsburger Seeblick 1, 10317 Berlin
Mobil: +49 1734685830
E-Mail: public-affairs@msg.group

Verantwortlich:
Dr. Jürgen Zehetmaier (Vorsitzender),
Michael Rasch,
Karsten Redenius,
Dr. Frank Schlottmann
Aufsichtsratsvorsitzender: Johann Zehetmaier