Themenreihe KI-native Entwicklung · Teil 4 von 5 · Teil 1: Was ist KI-native Softwareentwicklung? · Teil 2: Vom KI-Prototyp zur Produktion · Teil 3: KI-generierten Code prüfen · Teil 5: Wirtschaftlichkeit KI-nativer Entwicklung

Souveräne KI-Entwicklung: Datenklassen, Inferenz und Regulierung

Von **Thomas Faß** · 23. Oktober 2026 · aktualisiert 30. Oktober 2026

> **Ob eine KI-Entwicklung souverän ist, entscheidet die Architektur und weniger der Vertrag: Welche Daten welches Werkzeug sehen, wo die Modelle laufen, wer den Datenpfad kontrolliert und wer über Updates bestimmt. Diese Fragen gehören vor den ersten Build. Ein deutsches Rechenzentrum allein beantwortet sie nicht, denn der US CLOUD Act verpflichtet US-Anbieter zur Herausgabe von Daten unabhängig vom Speicherort. Die Nachfrage ist da: 73 Prozent der deutschen Unternehmen sorgen sich um die Abhängigkeit von den USA, 2024 waren es 38 Prozent (Bitkom, Oktober 2026).**

Das Wichtigste in Kürze

- Souveränität ist eine Architekturentscheidung. Sie entscheidet sich an vier Stellen: Modell, Inferenz, Datenpfad und Update-Takt. Eine Vertragsklausel ändert an keiner davon etwas.
- Der Speicherort allein schützt nicht. Der US CLOUD Act verpflichtet US-Anbieter, Daten auf Anordnung herauszugeben, auch wenn sie außerhalb der USA liegen.
- Die Rechtsgrundlage für US-Transfers ist in Bewegung. Das EU-US Data Privacy Framework gilt, ist aber vor dem EuGH angefochten. Ein US-Urteil vom Juni 2026 hat die Unabhängigkeit der FTC geschwächt, einer seiner Säulen.
- Nicht jedes Werkzeug sieht dieselben Daten. Wer den Bauprozess sauber vom Produktivsystem trennt und Daten in Klassen einteilt, kann im Bau leistungsfähige Modelle nutzen und die Produktivdaten trotzdem in Europa halten.
- Europäische Inferenz mit offenen Modellen ist verfügbar. Mehrere Anbieter betreiben sie in deutschen oder europäischen Rechenzentren. Offene Modelle sind aber nicht in jeder Aufgabe gleichwertig mit den stärksten proprietären.
- Die Regulierung hat sich verschoben. Die Hochrisiko-Pflichten des AI Act gelten nach dem Digital Omnibus ab 2. Dezember 2027. KI-Kompetenz und Transparenzpflichten gelten schon. NIS2 gilt in Deutschland seit dem 6. Dezember 2025.
- Die Nachfrage ist größer als die Umsetzung. In Österreich achten 64 Prozent der KMU bei der Cloud-Wahl auf den Speicherort, aber nur 49 Prozent wissen, wo ihre Daten tatsächlich liegen (Magenta Telekom, Oktober 2026).

Warum reicht eine DSGVO-Klausel nicht?

Viele Unternehmen regeln Souveränität im Vertrag: Auftragsverarbeitungsvertrag, Standardvertragsklauseln, Rechenzentrum in Frankfurt. Das ist notwendig. Es beantwortet aber nicht die Frage, wer im Ernstfall auf die Daten zugreifen kann.

Der CLOUD Act gilt unabhängig vom Speicherort. Das US-Gesetz von 2018 verpflichtet Anbieter, die US-Recht unterliegen, auf Anordnung Daten herauszugeben. Es spielt dabei keine Rolle, ob diese Daten in den USA oder anderswo liegen (18 U.S.C. § 2713). Ein deutsches Rechenzentrum eines US-Konzerns ändert daran nichts. Entscheidend ist, wer den Dienst betreibt und welchem Recht er unterliegt.

Schrems II hat die Prüfpflicht zum Unternehmen verlagert. Der Europäische Gerichtshof hat 2020 das Privacy Shield gekippt. Standardvertragsklauseln bleiben zulässig, doch Exporteur und Empfänger müssen im Einzelfall prüfen, ob das Schutzniveau im Zielland gleichwertig ist. Für die USA reichen die Klauseln allein in der Regel nicht (EuGH, C-311/18, 16.07.2020; Datenschutzkonferenz, 28.07.2020).

Das Data Privacy Framework gilt, steht aber unter Druck. Seit 2023 ist es die Grundlage für Datentransfers an zertifizierte US-Unternehmen. Das Gericht der EU hat eine Klage dagegen im September 2025 abgewiesen, der Kläger hat Rechtsmittel beim EuGH eingelegt (C-703/25 P, anhängig). Im Juni 2026 hat der US Supreme Court entschieden, dass der Präsident die Mitglieder der Handelsaufsicht FTC nach Belieben abberufen kann (Trump v. Slaughter). Die FTC ist eine zentrale Aufsichtsinstanz des Frameworks. Der Europäische Datenschutzausschuss hat die EU-Kommission im Juli 2026 gebeten zu prüfen, was das für das Framework bedeutet. Ausgesetzt ist es nicht.

Für die Architektur folgt daraus eine nüchterne Lehre. Wer heute ein KI-System baut, das Daten dauerhaft an einen US-Dienst gibt, baut auf einer Rechtsgrundlage, die sich während der Laufzeit ändern kann. Ein System, das den Anbieter wechseln kann, übersteht das. Eines, das es nicht kann, wird zum Notfallprojekt.

Was heißt Souveränität bei einem KI-System konkret?

Souveränität ist kein Etikett, sondern eine Reihe von Kontrollfragen. Für ein KI-System lassen sich vier Stellen unterscheiden, an denen sie sich entscheidet.

1. Das Modell. Welches Modell verarbeitet die Daten, und wer hat es trainiert? Bei einem offenen Modell liegen die Gewichte vor und lassen sich selbst betreiben. Bei einem proprietären Modell bleibt man an den Anbieter gebunden, auch wenn er das Modell in Europa anbietet.

2. Die Inferenz. Wo läuft das Modell, wenn es eine Anfrage beantwortet, und wer betreibt diese Umgebung? Ein offenes Modell bei einem US-Hyperscaler ist beim Betrieb so abhängig wie ein proprietäres. Ein offenes Modell bei einem europäischen Betreiber oder im eigenen Rechenzentrum ist es nicht.

3. Der Datenpfad. Welche Daten verlassen das Unternehmen, über welche Dienste, und was passiert dort mit ihnen? Dazu gehören nicht nur die Anfragen an das Modell. Auch Protokolle, Monitoring, Vektordatenbanken und Hilfsdienste wie E-Mail-Versand sind Teil des Pfads. Hier entstehen die meisten Lücken, weil sie beim Bauen niemand als Datenfluss betrachtet.

4. Der Update-Takt. Wer entscheidet, wann sich das Modell oder die Plattform ändert? Wenn ein Anbieter ein Modell abkündigt oder sein Verhalten ändert, verändert sich das eigene System mit. Wer Modelle selbst betreibt oder vertraglich feste Versionen nutzt, bestimmt den Takt selbst.

An diesen vier Stellen lässt sich auch ein Werbeversprechen prüfen. Ein Anbieter kündigt etwa für den 24. Oktober 2026 an, seine europäische Cloud mehrere Stunden ohne Verbindung zum globalen Netz zu betreiben (AWS Security Blog, 28.09.2026). Ein solcher Test kann zeigen, dass der Betrieb ohne Verbindung in die USA funktioniert. Die Frage, welchem Recht der Betreiber unterliegt, beantwortet er nicht. Beides gehört in die Bewertung, getrennt voneinander.

Die Bandbreite ist groß, und nicht jedes System braucht an jeder Stelle die strengste Lösung. Welche Stufe angemessen ist, entscheidet die Datenklasse.

Welches Werkzeug darf welche Daten sehen?

Der wichtigste Schritt ist eine Unterscheidung, die im Projektalltag oft untergeht. Ein KI-natives Vorhaben hat zwei Seiten: den **Bauprozess**, in dem Agenten Code schreiben, und das **Produktivsystem**, in dem die fertige Software echte Geschäftsdaten verarbeitet. Beide haben unterschiedliche Datenflüsse und brauchen unterschiedliche Regeln.

Im Bauprozess sieht das Werkzeug in der Regel Quellcode, Architekturentscheidungen und Testdaten. Im Produktivsystem sieht das eingebaute Modell Kundendaten, Verträge oder personenbezogene Daten. Wer beides gleich behandelt, verschenkt entweder Leistung im Bau oder Schutz im Betrieb.

Ein einfaches Raster mit vier Datenklassen hilft bei der Zuordnung:

Datenklasse	Beispiele	Bauprozess (Coding-Agenten)	Produktivsystem (eingebautes Modell)
1 Öffentlich	Dokumentation, offene Bibliotheken, Marketingtexte	jedes geprüfte Werkzeug	jedes geprüfte Modell
2 Intern	Quellcode, Architektur, Konfiguration ohne Geheimnisse	Werkzeuge mit Unternehmensvertrag, ohne Training auf Kundendaten	Modelle mit Unternehmensvertrag, bevorzugt in der EU
3 Vertraulich	Kundendaten, Preise, Verträge, Geschäftszahlen	nur synthetische oder anonymisierte Testdaten	Inferenz in der EU bei einem Betreiber unter EU-Recht
4 Streng vertraulich	Gesundheits- oder Personaldaten, Berufsgeheimnisse, Betriebsdaten kritischer Infrastruktur, Zugangsdaten	nie im Werkzeug, Geheimnisse außerhalb der Reichweite des Agenten	Inferenz im eigenen Rechenzentrum oder in einer dedizierten EU-Umgebung

Das Raster ist ein Ausgangspunkt, keine Rechtsauskunft. Welche Daten in welche Klasse gehören, hängt von Branche, Verträgen und Aufsicht ab. Zwei Regeln gelten aber fast immer.

Produktivdaten gehören nicht in den Bauprozess. Ein Agent braucht für die Entwicklung keine echten Kundendaten. Synthetische Testdaten oder anonymisierte Auszüge reichen. Damit bleibt der Bauprozess in Klasse 2, und das Unternehmen kann dort die leistungsfähigsten Werkzeuge einsetzen.

Zugangsdaten gehören nie in die Reichweite eines Agenten. Was ein Coding-Agent technisch erreichen kann, sollte man so behandeln, als könne es abfließen. Wie das bei Coding-Agenten konkret aussieht, beschreibt der Beitrag Wohin Ihr Code geht, wenn eine KI ihn schreibt. Die Prüfstufen aus Teil 3 gehören ebenfalls dazu: Ein Agent darf weder seine eigenen Prüfungen noch das Monitoring verändern.

Welche europäischen Optionen gibt es für die Inferenz?

Für Daten der Klassen 3 und 4 stellt sich die Frage, wo das Modell im Produktivsystem läuft. Das Angebot hat sich in den vergangenen anderthalb Jahren deutlich verbreitert. In alphabetischer Reihenfolge, ohne Wertung:

- **IONOS AI Model Hub:** Inferenz mit offenen Sprach- und Bildmodellen. Laut Dokumentation bleibt die Verarbeitung in Deutschland, Eingaben werden nicht zum Training genutzt.
- **OVHcloud AI Endpoints:** seit April 2025 verfügbar, mit mehr als 40 offenen Modellen. Betrieb aus europäischen Rechenzentren, das Angebot gibt es auch in anderen Regionen.
- **STACKIT AI Model Serving:** verwaltetes Hosting offener Sprachmodelle. STACKIT betreibt seine Regionen in Deutschland und Österreich.
- **T Cloud Public (ehemals Open Telekom Cloud), AI Foundation Services:** Schnittstelle für offene Modelle, gemeinsam genutzt oder dediziert, mit Hosting in der EU. Proprietäre Modelle im selben Angebot laufen bei anderen Anbietern, das gehört in die Prüfung.
- **Europäische Modellanbieter wie Mistral AI:** proprietäre und offene Modelle eines EU-Unternehmens, Daten standardmäßig in der EU. Je nach Funktion sind Übermittlungen an Unterauftragnehmer in Drittländern möglich.
- **Eigener Betrieb:** offene Modelle im eigenen Rechenzentrum oder in einer dedizierten Umgebung. Maximale Kontrolle über alle vier Stellen, aber eigener Aufwand für Hardware, Betrieb und Updates.

Bei der Auswahl helfen drei Kriterien. Erstens: Welche Modelle bietet der Dienst an, und reichen sie für die Aufgabe? Zweitens: Wo genau läuft die Inferenz, und gilt das für alle Funktionen oder nur für einen Teil? Drittens: Lässt sich das Modell bei einem anderen Betreiber in gleicher Form weiterbetreiben? Bei offenen Modellen ist das meist der Fall. Damit bleibt der Wechsel möglich, den der Data Act rechtlich erleichtert, aber nicht technisch herstellt. Was der Data Act dafür regelt, beschreibt der Fachartikel Cloud-Exit-Strategie.

Die ehrliche Einschränkung: Offene Modelle sind nicht in jeder Aufgabe gleichwertig mit den stärksten proprietären Modellen. Bei komplexem Schlussfolgern oder sehr langen Kontexten liegen proprietäre Modelle oft vorn. Souveränität kostet deshalb Leistung oder Aufwand. Das gehört offen abgewogen, Aufgabe für Aufgabe. Viele Anwendungsfälle im Mittelstand, etwa Klassifikation, Extraktion aus

Dokumenten oder Zusammenfassungen, laufen mit offenen Modellen gut. Ob das für den eigenen Fall gilt, zeigt ein Test mit echten Aufgaben, nicht ein Benchmark.

Was verlangen AI Act, NIS2 und BSI von einem KI-System?

Die Regulierung hat sich 2026 spürbar bewegt. Die wichtigsten Stände für Individualsoftware mit KI-Komponente:

AI Act. Die Verordnung (EU) 2024/1689 gilt gestaffelt. Seit dem 2. Februar 2025 gelten die Verbote und die Pflicht zur KI-Kompetenz nach Artikel 4: Unternehmen, die KI-Systeme anbieten oder einsetzen, müssen für ausreichende Kompetenz ihres Personals sorgen. Seit dem 2. August 2025 gelten die Pflichten für Anbieter von KI-Modellen mit allgemeinem Verwendungszweck. Seit dem 2. August 2026 gilt die Verordnung allgemein, einschließlich der Transparenzpflichten. Ein Chatbot muss sich etwa als KI zu erkennen geben.

Der Digital Omnibus zur KI, die Verordnung (EU) 2026/1744, hat die Hochrisiko-Pflichten verschoben. Für eigenständige Hochrisiko-Systeme nach Anhang III gelten sie ab dem 2. Dezember 2027, für KI in regulierten Produkten nach Anhang I ab dem 2. August 2028.

Für die meisten Individualentwicklungen im Mittelstand heißt das: Sie sind kein Hochrisiko-System. Das ändert sich, wenn die KI-Komponente in einen der Bereiche aus Anhang III fällt, etwa Personalauswahl, Kreditwürdigkeit oder Sicherheitskomponenten kritischer Infrastruktur. Dann sollte die Einstufung vor dem Bau stehen, weil sie Dokumentation, Risikomanagement und menschliche Aufsicht von Anfang an bestimmt.

NIS2. Das deutsche Umsetzungsgesetz gilt seit dem 6. Dezember 2025 ohne Übergangsfrist. Betroffen sind laut Fachpresse rund 29.500 Einrichtungen in 18 Sektoren, vorher waren es rund 4.500. Sie müssen sich beim BSI registrieren, Risikomanagement nachweisen und erhebliche Sicherheitsvorfälle melden: eine erste Meldung binnen 24 Stunden, eine Meldung binnen 72 Stunden, einen Abschlussbericht binnen eines Monats. Für ein KI-System heißt das: Es ist Teil der IT, die das Risikomanagement abdeckt. Wer einen externen Modelldienst einbindet, holt sich einen Lieferanten in die eigene Sicherheitskette.

BSI-Kriterien. Für Cloud-Dienste ist der C5-Katalog der Maßstab. Das BSI hat im April 2026 die Fassung C5:2026 veröffentlicht, Pflicht für Prüfungen ab dem 1. Juni 2027. Neu ist C3A, ein Kriterienkatalog für die Souveränität von Cloud-Diensten in sechs Dimensionen, vom Rechtlichen über die Daten bis zur Lieferkette. C3A ist kein Zertifikat, sondern ein Werkzeug für die Beschaffung, und setzt C5 voraus. Für KI-Cloud-Dienste gibt es zusätzlich den freiwilligen Katalog AIC4.

DSGVO. Sie gilt unverändert für jede Verarbeitung personenbezogener Daten, auch durch ein Modell. Rechtsgrundlage, Auftragsverarbeitung, Löschbarkeit und Auskunftsrechte müssen für den gesamten Datenpfad geklärt sein, einschließlich Protokollen und Vektordatenbanken.

Diese Übersicht ist eine fachliche Einordnung und ersetzt keine Rechtsberatung im Einzelfall.

Wie sieht das an einem konkreten Vorhaben aus?

Das Beispiel aus Teil 2 und Teil 3 geht weiter. Unternehmen, Personen und Details sind konstruiert, Messwerte enthält das Beispiel nicht.

Die Anforderung. Der Großhändler für Sanitärtechnik will seine Angebots-App erweitern. Kunden schicken Anfragen oft als Freitext oder als Foto einer handschriftlichen Liste. Ein Sprachmodell soll daraus einen Angebotsentwurf mit passenden Artikelnummern machen. Die Anfragen enthalten Namen, Adressen und Baustellenangaben, die Artikeldaten enthalten die Einkaufspreise.

Die Datenklassen. Die IT-Leitung ordnet zu: Kundenanfragen und Einkaufspreise sind Klasse 3. Der Code der App ist Klasse 2. Personenbezogene Daten besonderer Kategorien kommen nicht vor.

Der Bauprozess. Die Entwicklerin baut die Funktion mit einem Coding-Agenten unter Unternehmensvertrag. Der Agent bekommt den Code, die Schnittstellenbeschreibung und 200 synthetische Anfragen, die das Team aus echten Mustern nachgebaut hat. Echte Kundenanfragen sieht er nicht. Zugangsdaten zum Testsystem liegen in einem Tresor, auf den der Agent keinen Zugriff hat.

Das Produktivsystem. Für die Inferenz testet das Team zwei offene Modelle bei einem europäischen Betreiber mit 300 anonymisierten, echten Anfragen. Eines erkennt die Artikel zuverlässig genug, bei handschriftlichen Listen braucht es eine Nachkontrolle durch den Innendienst. Das Team entscheidet sich dafür und baut die Nachkontrolle als festen Schritt ein. Protokolle der Anfragen werden nach 30 Tagen gelöscht, die Vektordatenbank mit den Artikeldaten läuft in derselben Umgebung.

Die Wechselfähigkeit. Das Modell ist offen, der Aufruf läuft über eine einheitliche Schnittstelle. Ein Wechsel des Betreibers ist damit ein Konfigurationsschritt und ein erneuter Test, kein Umbau.

Das Ergebnis. Im Bau hat das Team die leistungsfähigsten Werkzeuge genutzt. Im Betrieb verlassen Kundendaten und Einkaufspreise die EU nicht. Der Preis dafür ist die Nachkontrolle bei handschriftlichen Listen. Das Team hat ihn bewusst bezahlt.

Checkliste für die Architekturentscheidung

Vor dem ersten Build sollten diese Fragen beantwortet sein:

1. Welche Datenklassen verarbeitet das System, im Bau und im Betrieb?
2. Welche Werkzeuge sehen im Bauprozess welche Daten, und gibt es für jedes einen Unternehmensvertrag?
3. Wo läuft die Inferenz im Produktivsystem, und welchem Recht unterliegt der Betreiber?
4. Welche Dienste gehören außer dem Modell zum Datenpfad: Protokolle, Monitoring, Vektordatenbank, E-Mail?
5. Wer entscheidet über Modellversionen und Updates?
6. Lässt sich das Modell bei einem anderen Betreiber weiterbetreiben, und wie lange würde der Wechsel dauern?

7. Fällt die KI-Komponente unter einen Bereich aus Anhang III des AI Act?
8. Ist das Unternehmen NIS2-pflichtig, und ist der Modelldienst als Lieferant im Risikomanagement erfasst?
9. Wo bleibt man bewusst hinter der strengsten Lösung zurück, und warum?

Die letzte Frage ist die wichtigste. Souveränität für jeden Datensatz in jeder Stufe ist teuer und selten nötig. Entscheidend ist, dass die Abwägung bewusst getroffen und dokumentiert ist.

Wie geht es in dieser Reihe weiter?

Dieser Artikel ist Teil 4 der fünfteiligen Themenreihe KI-native Entwicklung. Teil 1 hat geklärt, was KI-native Softwareentwicklung ist, Teil 2, was zwischen einem KI-Prototyp und dem Betrieb liegt, Teil 3, wie man KI-generierten Code prüft. Zum Abschluss:

- **Teil 5, 30. Oktober:** Wirtschaftlichkeit: wann sich KI-native Entwicklung lohnt und wann nicht.

Ihr nächster Schritt

Sie planen ein KI-Vorhaben mit sensiblen Daten und wollen die Architektur vor dem ersten Build klären? Sprechen Sie mit uns.

Häufige Fragen

Was bedeutet souveräne KI-Entwicklung?

Souveräne KI-Entwicklung heißt, dass ein Unternehmen über die entscheidenden Stellen seines KI-Systems selbst bestimmt: welches Modell die Daten verarbeitet, wo es läuft, welchen Weg die Daten nehmen und wer über Updates entscheidet. Diese Kontrolle entsteht durch die Architektur, nicht durch eine Vertragsklausel.

Reicht ein Rechenzentrum in Deutschland für Souveränität?

Nein. Der US CLOUD Act verpflichtet Anbieter, die US-Recht unterliegen, Daten auf Anordnung herauszugeben, auch wenn sie außerhalb der USA gespeichert sind. Entscheidend ist, wer den Dienst betreibt und welchem Recht er unterliegt, nicht nur der Standort des Rechenzentrums.

Darf man im regulierten Umfeld Claude, ChatGPT oder Copilot beim Entwickeln nutzen?

Oft ja, wenn Bauprozess und Produktivsystem getrennt sind. Im Bauprozess sieht ein Coding-Agent in der Regel Quellcode und Testdaten, keine Kundendaten. Mit Unternehmensvertrag, ohne Training auf den eigenen Daten, mit synthetischen Testdaten und ohne Zugangsdaten in Reichweite des Agenten ist das in vielen Branchen vertretbar. Ob es im Einzelfall zulässig ist, hängt von Branche, Verträgen und Aufsicht ab.

Sind offene Modelle sicherer als proprietäre?

Nicht automatisch. Ein offenes Modell lässt sich aber selbst oder bei einem europäischen Betreiber betreiben, damit liegen Inferenz, Datenpfad und Update-Takt in eigener Hand. Dafür sind offene Modelle nicht in jeder Aufgabe gleichwertig mit den stärksten proprietären. Ein Test mit echten Aufgaben zeigt, ob sie reichen.

Ab wann gelten die Hochrisiko-Pflichten des AI Act?

Nach dem Digital Omnibus zur KI, Verordnung (EU) 2026/1744, für eigenständige Hochrisiko-Systeme nach Anhang III ab dem 2. Dezember 2027 und für KI in regulierten Produkten nach Anhang I ab dem 2. August 2028. Die Pflicht zur KI-Kompetenz gilt seit dem 2. Februar 2025, die Transparenzpflichten seit dem 2. August 2026.

Was ändert der AI Act für eine Individualsoftware mit KI-Komponente?

In den meisten Fällen geht es um KI-Kompetenz der Beteiligten und um Transparenz, etwa den Hinweis, dass ein Chatbot eine KI ist. Strengere Pflichten zu Dokumentation, Risikomanagement und menschlicher Aufsicht kommen hinzu, wenn die Komponente in einen Hochrisiko-Bereich nach Anhang III fällt, zum Beispiel Personalauswahl oder Kreditwürdigkeit.

Ist das EU-US Data Privacy Framework noch gültig?

Ja, Stand Oktober 2026. Es ist aber vor dem EuGH angefochten (C-703/25 P), und nach einem Urteil des US Supreme Court zur Unabhängigkeit der FTC hat der Europäische Datenschutzausschuss die EU-Kommission um eine Prüfung gebeten. Wer US-Dienste einsetzt, sollte einen Wechsel einplanen können.

Was ist der BSI-Katalog C3A?

C3A ist ein Kriterienkatalog des BSI vom April 2026, mit dem sich die Souveränität von Cloud-Diensten in sechs Dimensionen bewerten lässt, darunter rechtliche, operative und Lieferketten-Aspekte. Er ist kein Zertifikat, sondern ein Werkzeug für die Beschaffung, und setzt den C5-Katalog voraus.

Gilt NIS2 auch für KI-Systeme?

NIS2 regelt nicht KI im Besonderen, sondern die Informationssicherheit betroffener Einrichtungen. Ein KI-System ist Teil dieser IT. Ist ein Unternehmen NIS2-pflichtig, gehören das System und ein eingebundener Modelldienst als Lieferant in das Risikomanagement und in die Meldeprozesse.

Was kostet souveräne KI im Vergleich?

Eine pauschale Zahl gibt es nicht. Souveränität kostet entweder Leistung, wenn ein offenes Modell schwächer ist als das stärkste proprietäre, oder Aufwand, wenn man Modelle selbst betreibt. Wer nach Datenklassen entscheidet, zahlt diesen Preis nur dort, wo die Daten es verlangen.

Thematische Vertiefungen

KI-native Softwareentwicklung für den Mittelstand

Souverän gemacht: unsere Sicht auf digitale Souveränität

Quellen

- Bitkom: „16 Monate könnten Deutschlands Unternehmen ohne Digitalimporte überleben“, Presseinformation, 19.10.2026 (602 Unternehmen ab 20 Beschäftigten). –

https://www.bitkom.org/Presse/Presseinformation/16-Monate-koennten-Deutschlands-Unternehmen-ohne-Digitalimporte-ueberleben

- Magenta Telekom: „KI, Cloud und Datensouveränität: Anspruch und Realität klaffen auseinander“, 21.10.2026 (800 Entscheider, KMU in Österreich). – https://newsroom.magenta.at/2026/10/21/ki-cloud-und-datensouveranitat-anspruch-und-realitat-klaffen-auseinander/
- US Congress: Clarifying Lawful Overseas Use of Data Act (CLOUD Act), S. 2383, 18 U.S.C. § 2713. – https://www.congress.gov/115/bills/s2383/BILLS-115s2383is.pdf
- Datenschutzkonferenz: Pressemitteilung zum Urteil des EuGH in der Rechtssache C-311/18 (Schrems II), 28.07.2020. – https://www.lda.bayern.de/media/pm/pm2020_7.pdf
- Digital Policy Alert: Rechtsmittel Latombe gegen das EU-US Data Privacy Framework, C-703/25 P. – https://digitalpolicyalert.org/event/35459
- Faegre Drinker: „Trump v. Slaughter: Implications of the US Supreme Court Ruling for EU-US Data Transfers and the Data Privacy Framework“, 17.09.2026. – https://www.faegredrinker.com/en/insights/publications/2026/9/trump-v-slaughter-implications-of-the-us-supreme-court-ruling-for-eu-us-data-transfers-and-the-data-privacy-framework
- Hunton: „EDPB Calls for Review of EU-U.S. Data Privacy Framework After U.S. Supreme Court Decision on FTC Independence“, 05.08.2026. – https://www.hunton.com/privacy-and-cybersecurity-law-blog/edpb-calls-for-review-of-eu-u-s-data-privacy-framework-after-u-s-supreme-court-decision-on-ftc-independence
- EU-Kommission: AI Act, Regulatory Framework und Anwendungsfristen. – https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
- Verordnung (EU) 2026/1744 (Digital Omnibus zur KI), Amtsblatt der EU. – https://eur-lex.europa.eu/eli/reg/2026/1744/oj?locale=de
- IHK Frankfurt am Main: „Der neue KI-Omnibus (Verordnung (EU) 2026/1744)“. – https://www.frankfurt-main.ihk.de/recht/uebersicht-alle-rechtsthemen/digitalisierung2/neue-ki-omnibus-verordnung-eu-2026-1744-7137020
- Reed Smith: „Finally: Germany enacts its NIS2 law“. – https://www.reedsmith.com/our-insights/blogs/technology-law-dispatch/102lxfr/finally-germany-enacts-its-nis2-law/
- SECJUR: NIS2UmsuCG, Pflichten und Meldefristen. – https://www.secjur.com/blog/nis2umsucg
- Rödl & Partner: „C5:2026“, 08.04.2026. – https://www.roedl.com/insights/c5-2026-cloud-compliance/
- Vergabeblog: „BSI veröffentlicht C3A“, 30.04.2026. – https://vergabeblog.de/2026-04-30/bsi-veroeffentlicht-c3a-neuer-massstab-fuer-digitale-souveraenitaet-in-der-cloud-beschaffung/
- BSI: AIC4, Kriterienkatalog für KI-Cloud-Dienste. – https://www.bsi.bund.de/EN/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/AIC4/aic4.html
- AWS Security Blog: „AWS European Sovereign Cloud: Demonstrating an independent operation“, 28.09.2026. – https://aws.amazon.com/blogs/security/aws-european-sovereign-cloud-demonstrating-an-independent-operation/
- IONOS: AI Model Hub, Dokumentation. – https://docs.ionos.com/cloud/ai/ai-model-hub/ai-model-hub

- OVHcloud: AI Endpoints, Pressemitteilung, 29.04.2025. – https://corporate.ovhcloud.com/en/newsroom/news/ai-endpoints/
- STACKIT: AI Model Serving und Regionen, Dokumentation. – https://docs.stackit.cloud/products/data-and-ai/ai-model-serving/ · Regionen – https://docs.stackit.cloud/platform/regions/
- T Cloud Public: AI Foundation Services. – https://www.t-cloud-public.com/en/solutions/use-cases/artificial-intelligence/ai-foundation-services
- Mistral AI: „Where do you store my data?“, Hilfe-Center. – https://help.mistral.ai/en/articles/347629-where-do-you-store-my-data-or-my-organization-s-data

---

Über den Autor

Thomas Faß ist Gründer und CEO von InnoShore, einem IT-Dienstleister für den DACH-Mittelstand, und leitet die Regionalgruppe NRW im KI Bundesverband. *Offenlegung: InnoShore ist IONOS CLOUD Channel Partner und STACKIT Reseller-Partner für Cloud-Infrastruktur und KI-Inferenz.*