KI-WELTREISE

Factbook Staffel 9

Feuer unter dem Eis: Autarke KI-Systeme ohne Cloud-Zwang

Robert Hortschitz
17.5.2026

Inhalt

IDEE, TEXT, KONZEPT & LERNAUFBEREITUNG: BIRGIT POHN & ROBERT HORTSCHITZ; OPTIMIERT UND UNTERSTÜTZT MIT DEN KI SYSTEMEN CHATGPT, COPILOT, GEMINI, MISTRAL, NOTEBOOKLM & CLAUDE; EINE PRODUKTION DER MOGI BUSINESS CREATION COMANY GMBH & STRO GMBH; COPYRIGHT 2026

Executive Überblick

Staffel 9 behandelt eine der wichtigsten Architekturfragen der nächsten Jahre: Muss KI immer in der Cloud laufen, oder können Unternehmen leistungsfähige KI auch lokal, am Rand des Netzwerks oder direkt an der Maschine betreiben?

Die Staffel nutzt Island als durchgehende Metapher. Island ist geologisch aktiv, wetterhart, teilweise isoliert und trotzdem hoch funktional. Die Energie kommt aus dem eigenen Boden. Genau dieses Bild trägt die Staffel: KI muss nicht immer aus entfernten Rechenzentren kommen. In vielen Fällen ist es stabiler, schneller und datenschutzfreundlicher, wenn Verarbeitung dort stattfindet, wo die Daten entstehen.

Für Österreich und Deutschland ist das Thema besonders relevant, weil viele Unternehmen mit Datenschutz, Betriebsgeheimnissen, industriellen Echtzeitprozessen, EU AI Act, DSGVO, NIS2, DORA und wachsender Cloud-Abhängigkeit umgehen müssen. Die richtige Antwort ist nicht „Cloud oder lokal“, sondern eine saubere hybride Architektur.

1. Edge AI

Einfach erklärt:
Edge AI bedeutet, dass KI nicht in einem entfernten Cloud-Rechenzentrum läuft, sondern nahe am Ort, an dem Daten entstehen. Das kann eine Maschine, ein Fahrzeug, eine Kamera, ein lokaler Server, ein Industrie-PC oder ein Standort-Rechenzentrum sein.

Bezug zur Staffel:
Die isolierte Fabrik in Island zeigt Edge AI praktisch: Kameras prüfen Bauteile, Sensoren messen Druck und Temperatur, lokale Systeme entscheiden sofort. Die Daten müssen nicht zuerst über das Internet verschickt werden.

Warum relevant:
Edge AI reduziert Latenz, senkt Datenübertragung, erhöht Ausfallsicherheit und hilft, sensible Daten lokal zu halten. Red Hat beschreibt Edge AI als Kombination von KI und Edge Computing, bei der Daten nahe an einem physischen Ort gesammelt und verarbeitet werden.

Praxisbeispiele:
Qualitätskontrolle in der Produktion, Predictive Maintenance, lokale Videoanalyse, autonome Geräte, medizinische Geräte, Logistik-Hubs, Smart Buildings.

AI Act / DSGVO:
Edge AI kann DSGVO-Risiken reduzieren, weil weniger personenbezogene Daten übertragen werden. Es löst die DSGVO aber nicht automatisch. Rechtsgrundlage, Zweckbindung, Datenminimierung, Zugriffsschutz und Protokollierung bleiben notwendig. Beim AI Act zählt nicht der Ort der Verarbeitung allein, sondern Zweck, Risiko und Einsatzbereich des Systems. Der AI Act folgt einem risikobasierten Ansatz.

2. Local LLM

Einfach erklärt:
Ein Local LLM ist ein Sprachmodell, das auf eigener Infrastruktur läuft. Also nicht über eine externe API, sondern auf einem lokalen Server, einer Workstation, einem Edge-Gerät oder einem privaten Rechenzentrum.

Bezug zur Staffel:
Das „Rechenzentrum im Vulkan“ steht für lokale Sprachmodelle. Daten bleiben im eigenen Umfeld. Modelle werden kontrolliert aktualisiert. Antworten bleiben reproduzierbarer als bei ständig wechselnden Cloud-Diensten.

Warum relevant:
Lokale LLMs sind besonders interessant für Unternehmen mit vertraulichen Dokumenten, Betriebswissen, Kundendaten, IP-Schutz, Forschungsdaten oder regulierten Prozessen.

Grenzen:
Lokale Modelle sind oft kleiner als große Cloud-Modelle. Sie brauchen passende Hardware, Betrieb, Monitoring, Patchmanagement, Modellpflege und klare Verantwortlichkeiten.

AI Act / DSGVO:
Wenn ein Unternehmen ein lokales Modell nur intern einsetzt, bleibt es trotzdem Verantwortlicher oder Auftragsverarbeiter im Sinne der DSGVO, je nach Verarbeitungskontext. Beim AI Act kann das Unternehmen Betreiber eines KI-Systems sein. Werden eigene Modelle bereitgestellt oder wesentlich verändert und am Markt angeboten, können zusätzliche Pflichten entstehen. Für General Purpose AI gelten seit 2. August 2025 eigene Pflichten für Anbieter solcher Modelle.

3. On-Premise AI

Einfach erklärt:
On-Premise AI bedeutet, dass KI-Systeme im eigenen Rechenzentrum oder im eigenen Standort betrieben werden. Die Infrastruktur gehört oder unterliegt direkt der Kontrolle des Unternehmens.

Abgrenzung zu Edge AI:
On-Premise ist meist zentral im Unternehmen. Edge ist näher an Maschinen, Sensoren oder Standorten. Ein Unternehmen kann beides kombinieren.

Praxisnutzen:
On-Premise AI ist sinnvoll bei sensiblen Daten, bestehenden Rechenzentrumsstrukturen, niedriger Toleranz gegenüber externen Abhängigkeiten oder klaren Compliance-Anforderungen.

Risiko:
Eigene Infrastruktur bedeutet eigene Verantwortung: Hardware, Strom, Kühlung, Sicherheit, Updates, Backup, Monitoring, Modellversionierung.

4. Hybrid AI

Einfach erklärt:
Hybrid AI kombiniert lokale Verarbeitung, Edge-Systeme und Cloud-Dienste. Nicht alles läuft lokal, nicht alles läuft in der Cloud.

Bezug zur Staffel:
Kapitel 5, „Der hybride Pfad“, ist die wichtigste Architekturlektion. Lokal wird gehandelt, zentral wird analysiert. Edge verarbeitet Echtzeitdaten, die Cloud oder ein zentrales Rechenzentrum übernimmt Training, Reporting, Flottenmanagement oder langfristige Optimierung.

Gute Aufgabenteilung:
Lokal: Echtzeit, sensible Daten, Ausfallsicherheit, Vorverarbeitung.
Zentral: Training, Modellmanagement, Langzeitanalyse, übergreifende Optimierung, Reporting.

AI Act / DSGVO:
Hybride Architekturen müssen sauber dokumentiert werden. Wichtig sind Datenflüsse, Rollen, Rechtsgrundlagen, Speicherorte, Zugriffskonzepte und Auftragsverarbeitung. Für Hochrisiko-Systeme nach AI Act sind Dokumentation, Risikomanagement, menschliche Aufsicht und Datenqualität besonders relevant.

5. Cloud Dependency

Einfach erklärt:
Cloud Dependency ist die Abhängigkeit eines Systems von externen Cloud-Diensten. Wenn Verbindung, Dienst, Lizenz, Region oder Anbieter nicht verfügbar sind, kann das eigene System eingeschränkt sein.

Bezug zur Staffel:
Der Sturm in Keflavík zeigt genau dieses Problem. Das Navigationssystem funktioniert nur eingeschränkt, weil es ständig Daten nachladen muss.

Typische Abhängigkeiten:
APIs, Identitätsdienste, Modellendpunkte, Cloud-Speicher, Lizenzserver, Monitoring, zentrale Datenbanken.

Praxisfrage:
Welche Funktionen müssen auch ohne Cloud-Verbindung weiterlaufen?

6. Latenz

Einfach erklärt:
Latenz ist die Verzögerung zwischen Anfrage und Antwort. In klassischen Büroprozessen ist sie oft nur lästig. In Echtzeitprozessen kann sie kritisch sein.

Bezug zur Staffel:
In der isolierten Fabrik muss ein fehlerhaftes Bauteil sofort erkannt werden. Wenn Bilddaten erst in eine entfernte Cloud übertragen werden müssen, kann die Entscheidung zu spät kommen.

Praxisbeispiele:
Maschinensteuerung, Robotik, Qualitätsprüfung, medizinische Geräte, Sicherheitssysteme, autonome Fahrzeuge.

7. Datenhoheit

Einfach erklärt:
Datenhoheit bedeutet, dass ein Unternehmen Kontrolle darüber behält, wo Daten liegen, wer darauf zugreift, wie sie verarbeitet werden und ob sie weitergegeben werden.

Bezug zur Staffel:
Im Rechenzentrum im Vulkan bleiben interne Daten lokal. Das ist das Gegenbild zum dauerhaften Senden sensibler Daten an externe Systeme.

DSGVO-Relevanz:
Datenhoheit ist kein Ersatz für Datenschutz. Sie unterstützt aber Datenschutzprinzipien wie Datenminimierung, Zugriffsbeschränkung und Zweckbindung. Die EDPS-Leitlinien zu generativer KI betonen praktische Maßnahmen zur Einhaltung des Datenschutzes bei der Nutzung generativer KI-Systeme.

8. Data Sovereignty

Einfach erklärt:
Data Sovereignty bedeutet, dass Daten den rechtlichen und organisatorischen Anforderungen eines bestimmten Landes, einer Region oder Organisation unterliegen.

Für Österreich und Deutschland:
Relevant bei öffentlichen Auftraggebern, kritischer Infrastruktur, Gesundheitsdaten, Forschung, Industrie-Know-how, Behörden und Finanzdienstleistungen.

Wichtig:
Ein Rechenzentrum in Europa hilft, löst aber nicht alle Fragen. Entscheidend sind auch Anbieterstruktur, Zugriffsmöglichkeiten, Unterauftragsverarbeiter, Supportzugriffe und Vertragslage.

9. Air-Gapped AI

Einfach erklärt:
Air-Gapped AI läuft in einer Umgebung, die vom Internet oder externen Netzen getrennt ist. Solche Systeme werden für besonders sensible oder kritische Anwendungen verwendet.

Services:
Google Distributed Cloud bietet eine air-gapped Variante für sichere On-Premise-Umgebungen und beschreibt Vertex AI auf Distributed Cloud als Lösung für Organisationen mit hohen Anforderungen an Datensouveränität, Sicherheit und Datenschutz.

Risiko:
Air-gapped ist sicherer gegen externe Abhängigkeiten, aber schwieriger zu betreiben. Updates, Modellpflege, Schwachstellenmanagement und Logging müssen geplant erfolgen.

10. Model Inference

Einfach erklärt:
Inference ist die Nutzung eines trainierten Modells. Das Modell erzeugt eine Antwort, erkennt ein Objekt, klassifiziert ein Bild oder trifft eine Vorhersage.

Bezug zur Staffel:
Die Fabrik führt lokale Inference aus: Kameradaten werden direkt bewertet.

Praxis:
Viele Edge-Szenarien trainieren Modelle zentral, führen die Inference aber lokal aus. Red Hat beschreibt dieses Muster mit OpenShift AI: zentral trainieren oder tunen, dann an Edge-Standorten für Inference bereitstellen.

11. Model Training

Einfach erklärt:
Training ist der Prozess, bei dem ein Modell aus Daten lernt. Training ist meist rechenintensiver als Inference.

Praxis:
Für viele Unternehmen ist es sinnvoller, bestehende Modelle zu verwenden, anzupassen oder mit Retrieval zu kombinieren, statt eigene große Modelle von Grund auf zu trainieren.

AI Act:
Wer Modelle entwickelt, wesentlich verändert oder als General Purpose AI bereitstellt, kann in Pflichten eines Anbieters fallen. Die EU-Kommission hat Leitlinien zu Pflichten für Anbieter von General-Purpose-AI-Modellen veröffentlicht.

12. Fine-Tuning

Einfach erklärt:
Fine-Tuning bedeutet, ein bestehendes Modell mit zusätzlichen Daten auf einen bestimmten Zweck anzupassen.

Nutzen:
Bessere Domänensprache, konsistentere Antworten, spezielle Aufgaben.

Risiken:
Training mit personenbezogenen Daten kann DSGVO-pflichtig sein. Außerdem können vertrauliche Informationen unbeabsichtigt im Modellverhalten auftauchen, wenn Daten ungeeignet verwendet werden.

13. RAG, Retrieval-Augmented Generation

Einfach erklärt:
RAG verbindet ein Sprachmodell mit einer Wissensbasis. Das Modell beantwortet Fragen nicht nur aus seinem internen Musterwissen, sondern holt relevante Informationen aus Dokumenten, Datenbanken oder Suchsystemen.

Warum wichtig:
Für Unternehmen ist RAG oft sinnvoller als Fine-Tuning. Dokumente bleiben besser aktualisierbar, Quellen können nachvollzogen werden, Berechtigungen lassen sich sauberer abbilden.

Services:
NVIDIA NeMo Retriever ist ein Beispiel für Enterprise-RAG-Komponenten mit Retrieval, Vektorsuche und LLM-Unterstützung.

DSGVO / AI Act:
RAG braucht Zugriffskontrolle. Ein Modell darf keine Dokumente ausgeben, auf die der Nutzer keinen Zugriff hat. Für AI-Act-relevante Anwendungen sind Nachvollziehbarkeit und Dokumentation wesentlich.

14. Vector Database

Einfach erklärt:
Eine Vektordatenbank speichert Inhalte nicht nur als Text, sondern als numerische Repräsentation ihrer Bedeutung. Dadurch lassen sich ähnliche Inhalte finden, auch wenn andere Wörter verwendet werden.

Praxis:
Wichtig für RAG, semantische Suche, Wissensassistenten und Dokumentenabfragen.

DSGVO:
Auch Vektoren können personenbezogene Informationen enthalten oder Rückschlüsse ermöglichen. Daher gehören sie in Datenschutzkonzepte, Löschkonzepte und Zugriffskontrollen.

15. Quantisierung

Einfach erklärt:
Quantisierung reduziert die Genauigkeit numerischer Modellwerte, damit Modelle kleiner, schneller und ressourcenschonender laufen.

Praxisnutzen:
Damit können Modelle auf kleineren GPUs, CPUs oder Edge-Geräten laufen.

Tooling:
Intel OpenVINO unterstützt Optimierung und Bereitstellung von KI-Modellen in Cloud, On-Premise und Edge, inklusive generativer Modelle.

16. Small Language Model, SLM

Einfach erklärt:
Ein SLM ist ein kleineres Sprachmodell. Es kann weniger allgemein sein als ein großes Modell, ist aber oft schneller, günstiger und leichter lokal betreibbar.

Bezug zur Staffel:
Für lokale Anwendungen in der Fabrik oder im Vulkan-Rechenzentrum sind SLMs oft realistischer als riesige Modelle.

Services:
Microsoft Foundry Local zielt auf On-Device- und lokale KI-Anwendungen und beschreibt den sicheren Entwurf, die Anpassung und Verwaltung von KI-Anwendungen und Agenten auf Geräten.

17. GPU

Einfach erklärt:
Eine GPU ist ein Prozessor, der viele Berechnungen parallel ausführen kann. Für KI-Inference und Training ist sie häufig wesentlich effizienter als klassische CPUs.

Praxis:
Lokale LLMs brauchen je nach Modellgröße ausreichend GPU-Speicher. Für kleinere Modelle reichen teilweise Workstations oder Edge-Geräte, für größere Systeme braucht es Server-Hardware.

18. NPU

Einfach erklärt:
Eine NPU ist ein spezieller Prozessor für KI-Berechnungen, häufig in modernen PCs, Smartphones oder Edge-Geräten.

Nutzen:
Lokale KI-Funktionen können energieeffizienter ausgeführt werden.

19. Edge Device

Einfach erklärt:
Ein Edge Device ist ein Gerät am Rand des Netzwerks. Dazu zählen Industrie-PCs, Kameras, Sensor-Gateways, Fahrzeuge, medizinische Geräte oder lokale Server.

Praxis:
Edge Devices müssen robust, updatefähig, sicher und verwaltbar sein. Ohne Gerätemanagement entsteht Wildwuchs.

20. Fleet Management

Einfach erklärt:
Fleet Management bedeutet, viele Edge-Geräte zentral zu verwalten. Updates, Konfiguration, Zertifikate, Monitoring und Modellversionen müssen kontrolliert ausgerollt werden.

Services:
AWS IoT Greengrass ermöglicht lokale Ausführung von Workloads auf Edge-Geräten. AWS weist darauf hin, dass Greengrass V1 am 1. Juni 2026 aus dem Support läuft und Greengrass V2 genutzt werden soll.

21. Model Drift

Einfach erklärt:
Model Drift bedeutet, dass ein Modell mit der Zeit schlechter wird, weil sich Realität, Daten oder Prozesse verändern.

Beispiel:
Eine Qualitätskontrolle wurde auf alte Bauteile trainiert. Neue Materialien oder Lieferanten verändern die Optik. Das Modell erkennt Fehler schlechter.

Pflicht in der Praxis:
Monitoring, Testdaten, regelmäßige Bewertung und dokumentierte Modellfreigaben.

22. MLOps

Einfach erklärt:
MLOps ist der geordnete Betrieb von KI-Modellen. Es umfasst Entwicklung, Test, Deployment, Monitoring, Versionierung, Rollback und Dokumentation.

Warum wichtig:
Ohne MLOps werden lokale und hybride KI-Systeme schwer kontrollierbar.

AI Act:
Für regulierte Systeme ist MLOps ein praktischer Baustein, um technische Dokumentation, Risikomanagement, Monitoring und Nachvollziehbarkeit umzusetzen.

23. Observability

Einfach erklärt:
Observability bedeutet, dass man versteht, was ein System tut. Logs, Metriken, Traces, Modellantworten, Fehlerraten und Latenzen werden sichtbar gemacht.

Praxis:
Bei lokalen KI-Systemen ist Observability besonders wichtig, weil Probleme nicht automatisch beim Cloud-Anbieter sichtbar sind.

24. Human Oversight

Einfach erklärt:
Human Oversight bedeutet menschliche Aufsicht. Menschen bleiben in relevanten Entscheidungen eingebunden.

AI Act:
Der AI Act sieht für bestimmte Risikoklassen Anforderungen an menschliche Aufsicht vor. Besonders bei Hochrisiko-KI ist das zentral.

Praxis:
Nicht jede KI-Antwort darf automatisch eine Entscheidung auslösen. Es braucht klare Freigabepunkte.

25. Guardrails

Einfach erklärt:
Guardrails sind technische und organisatorische Leitplanken. Sie begrenzen, was ein KI-System tun darf.

Beispiele:
Keine Ausgabe vertraulicher Daten, keine Aktionen ohne Freigabe, Filter für sensible Inhalte, Rollenmodelle, Zugriffskontrolle, Logging.

26. Confidential Computing

Einfach erklärt:
Confidential Computing schützt Daten während der Verarbeitung, nicht nur beim Speichern oder Übertragen.

Praxis:
Relevant bei sensiblen Cloud- oder Hybrid-Szenarien. Für lokale KI kann es zusätzliche Sicherheit schaffen, ist aber kein Ersatz für Datenschutzkonzepte.

Überblick: Services und Derivate für Österreich und Deutschland

Microsoft

Relevante Bausteine:
Microsoft Foundry, Foundry Local, Azure AI Services, Azure Local, Microsoft 365 Copilot, Copilot Studio.

Für Staffel 9 besonders relevant:
Foundry Local, weil Microsoft hier lokale und On-Device-KI-Anwendungen adressiert. Die Dokumentation beschreibt Foundry Local als Möglichkeit, KI-Anwendungen und Agenten auf Geräten sicher zu entwickeln, anzupassen und zu verwalten.

Stärken:
Integration in Microsoft-Ökosystem, Identität über Entra ID, Governance, Enterprise-Verträge.

Risiken:
Lizenz- und Produktänderungen, Abhängigkeit von Microsoft-Architektur, Datenzugriffe über Graph und M365-Berechtigungen.

DSGVO / AI Act:
Für Österreich und Deutschland sind Auftragsverarbeitungsverträge, Datenregion, Berechtigungsmodell, Purview, Logging und Copilot-Governance wesentlich. Bei produktiven KI-Anwendungen ist zu prüfen, ob es sich um normale Assistenzsysteme oder um AI-Act-relevante Systeme handelt.

NVIDIA

Relevante Bausteine:
NVIDIA AI Enterprise, Jetson, IGX, RTX PRO Server, NeMo, NIM, Triton Inference Server.

Für Staffel 9 besonders relevant:
NVIDIA positioniert Edge AI für Echtzeitentscheidungen und bietet Hardware- und Softwarekomponenten für Enterprise Edge, Embedded Edge und Industrial Edge.

Stärken:
GPU-Ökosystem, Edge-Hardware, industrielle Szenarien, KI-Inferenz, RAG-Komponenten.

Risiken:
Kosten, Hardwarebindung, Energiebedarf, Verfügbarkeit, Betriebs-Know-how.

DSGVO / AI Act:
NVIDIA liefert häufig Infrastruktur, nicht die fachliche Verantwortung. Unternehmen bleiben für Daten, Zweck, Modellverhalten und Governance verantwortlich.

Google Cloud

Relevante Bausteine:
Google Distributed Cloud, Vertex AI auf Distributed Cloud, Gemini für On-Premise-Szenarien.

Für Staffel 9 besonders relevant:
Google Distributed Cloud wird als vollständig verwaltete Software- und Hardwarelösung für Rechenzentren und Edge-Standorte beschrieben, unter anderem für regulatorische Anforderungen, lokale Datenverarbeitung, Ausfallsicherheit und niedrige Latenz.

Stärken:
Starke KI-Plattform, verteilte Cloud-Optionen, air-gapped Szenarien.

Risiken:
Enterprise-Komplexität, Vertrags- und Betriebsmodell, Anbieterabhängigkeit.

DSGVO / AI Act:
Für öffentliche Auftraggeber und regulierte Branchen sind Datenstandort, Zugriffsmöglichkeiten, Unterauftragsverarbeiter und Auditfähigkeit besonders wichtig.

AWS

Relevante Bausteine:
AWS IoT Greengrass V2, AWS IoT Core, Lambda am Edge, Container Workloads, SageMaker für zentrale ML-Prozesse.

Für Staffel 9 besonders relevant:
AWS IoT Greengrass unterstützt lokale Ausführung von Workloads auf Edge-Geräten. Wichtig: Greengrass V1 läuft am 1. Juni 2026 aus dem Support, Migration auf V2 ist notwendig.

Stärken:
IoT-Integration, Skalierung, Cloud-to-Edge-Management.

Risiken:
Greengrass V1 End-of-Support, Cloud-Integration muss bewusst geplant werden, Kosten- und Betriebsmodell.

DSGVO / AI Act:
Edge-Datenflüsse müssen dokumentiert werden. Bei personenbezogenen Sensordaten oder Videoanalyse sind Datenschutz-Folgenabschätzung, Zweckbindung und technische Schutzmaßnahmen zu prüfen.

Red Hat / OpenShift AI

Relevante Bausteine:
Red Hat OpenShift AI, OpenShift, Kubernetes, Hybrid Cloud, Edge Deployment.

Für Staffel 9 besonders relevant:
OpenShift AI verwaltet den Lebenszyklus von predictive und generative AI-Modellen über hybride Cloud-Umgebungen hinweg.

Stärken:
Hybrid- und Open-Source-Nähe, Kubernetes-Betriebsmodell, passend für Unternehmen mit OpenShift-Strategie.

Risiken:
Betriebsaufwand, Plattformkompetenz, klare MLOps-Prozesse notwendig.

DSGVO / AI Act:
Gut geeignet für kontrollierte Betriebsmodelle, wenn Logging, Zugriff, Modellfreigaben und Dokumentation sauber umgesetzt werden.

Intel / OpenVINO

Relevante Bausteine:
OpenVINO, Intel AI Edge Portfolio, CPU/GPU/NPU-Optimierung.

Für Staffel 9 besonders relevant:
OpenVINO ist ein Open-Source-Toolkit für performante KI-Bereitstellung in Cloud, On-Premise und Edge.

Stärken:
Gute Option für kosteneffiziente Inference auf Intel-Hardware, Edge-PCs, Industrie-PCs, AI-PCs.

Risiken:
Nicht jedes Modell ist gleich gut optimierbar. Performance muss praktisch getestet werden.

DSGVO / AI Act:
OpenVINO ist Werkzeugschicht. Rechtliche Verantwortung liegt beim Betreiber und Einsatzzweck.

Open-Source-Derivate

Relevante Bausteine:
Ollama, llama.cpp, vLLM, LocalAI, Hugging Face Transformers, LangChain, LlamaIndex, Haystack, Milvus, Qdrant, Weaviate, Chroma, Kubernetes, K3s.

Stärken:
Hohe Flexibilität, lokale Kontrolle, geringe Einstiegshürden, starke Entwicklercommunity.

Risiken:
Lizenzfragen, Modellherkunft, Security Updates, fehlender Enterprise-Support, Auditfähigkeit.

DSGVO / AI Act:
Open Source ist nicht automatisch frei von Pflichten. Modelllizenz, Trainingsdaten, personenbezogene Daten, Dokumentation und Risikoanalyse müssen geprüft werden.

Auswirkungen auf EU AI Act

Für Staffel 9 sind vor allem drei Rollen wichtig:

1. Anbieter eines KI-Systems
Wer ein KI-System entwickelt oder unter eigenem Namen bereitstellt, kann Anbieterpflichten haben.

2. Betreiber eines KI-Systems
Wer ein KI-System beruflich nutzt, kann Betreiberpflichten haben.

3. Anbieter eines General-Purpose-AI-Modells
Wer ein allgemeines Modell am Markt bereitstellt, kann GPAI-Pflichten haben. Die EU-Kommission nennt für GPAI-Anbieter Pflichten ab 2. August 2025.

Für Unternehmen in Österreich und Deutschland bedeutet das: Der lokale Betrieb eines Modells reduziert nicht automatisch AI-Act-Pflichten. Entscheidend ist, wofür das System eingesetzt wird. Ein lokaler Chatbot für interne Wissenssuche ist anders zu bewerten als ein System für Bewerberauswahl, Kreditentscheidung, medizinische Triage oder kritische Infrastruktur.

Auswirkungen auf DSGVO

Für DSGVO sind bei Staffel 9 besonders wichtig:

- Werden personenbezogene Daten verarbeitet?
- Bleiben Daten lokal oder gehen sie an Dritte?
- Gibt es eine Rechtsgrundlage?
- Werden Daten minimiert?
- Gibt es Löschkonzepte?
- Sind Betroffenenrechte umsetzbar?
- Sind Auftragsverarbeiter sauber geregelt?
- Gibt es Zugriffskontrolle und Protokollierung?
- Ist eine Datenschutz-Folgenabschätzung notwendig?

Die EDPB hat 2024 eine Stellungnahme zu Datenschutzaspekten bei KI-Modellen veröffentlicht, unter anderem zu Rechtsgrundlage und personenbezogenen Daten im Kontext von KI-Modellen.

Lokale KI kann DSGVO-freundlicher sein, wenn sie Datenübertragung reduziert. Sie kann aber auch neue Risiken schaffen, wenn lokale Systeme schlecht abgesichert, nicht dokumentiert oder ohne Löschkonzept betrieben werden.

Schlussgedanke

Staffel 9 macht einen wichtigen Punkt sichtbar: Die Zukunft produktiver KI liegt nicht in einer einzigen Betriebsform. Sie liegt in der sauberen Entscheidung, welche Verarbeitung wohin gehört.

Cloud bleibt wichtig.
Edge wird wichtiger.
Lokale LLMs werden realistischer.
Hybride Architekturen werden zum Normalfall.

Für Österreich und Deutschland ist das keine technische Modefrage, sondern eine Frage von Datenschutz, Resilienz, Betriebsfähigkeit und Verantwortung.