Boost AI Newsletter

MONTHLY BACKGROUND SERIES

Interreg North-West Europe
Co-funded by the European Union
Boost AI

Boost AI – Monthly Background Newsletter

September 2026 Edition

AI Sovereignty for SMEs: Where Should Your AI Run?

Why AI Sovereignty Matters

AI sovereignty sounds like a strategic topic for governments and large corporations, but it is becoming a very practical question for SMEs. Where your AI runs affects cost, availability, performance, data protection, trade secrets, vendor dependence, and how easily you can change technology later.

Europe is also putting more emphasis on technological sovereignty in AI, cloud, semiconductors, and open source. For an SME, however, sovereignty does not mean avoiding every non-European component. It means understanding which parts of your AI stack you control - and choosing the level of control that matches the sensitivity and importance of the use case.

Sovereignty is more than server location

Data & trade secrets
Where do prompts, documents, logs, embeddings, and model outputs go - and who can technically access them?

Legal jurisdiction
Which company is your contractual provider, which laws can apply to it, and are international data transfers involved?

Operational control
Can you keep working if the internet, an API, a provider, or a specific model becomes unavailable?

Technology & lock-in
Can you move your models, data, APIs, and workflows elsewhere without rebuilding everything?

Economics & skills
Do you prefer capital expenditure and in-house operation, or variable cloud cost and managed services?

Even local AI still depends on global hardware, software, models, and supply chains. The goal is not absolute independence, but deliberate control of the dependencies that matter.

From Your Laptop to the Global Cloud

There is no single "sovereign" deployment model. The following spectrum shows seven common ways an SME can run AI - from fully local inference to globally operated AI services - and the main trade-offs at each level.

Seven Deployment Levels - Main Advantages and Trade-offs

The levels below are not a compliance ranking. They show how operational control, elasticity, legal exposure, cost structure, and technical dependence typically change as more of the stack is operated by an external provider.

More direct control, offline capability, fixed capacity ← → **More elasticity, managed services, newest models**

Sovereignty is a spectrum - not a yes/no label

Level	Typical setup / examples	Main advantages	Main trade-offs
1 Local AI on a device or workstation	MacBook with Apple silicon, GPU workstation, NVIDIA DGX Spark, dedicated AI box	Maximum direct control over data and trade secrets; can work offline; very low latency; no per-token or per-request fee after the hardware is bought.	Limited by local memory and compute; larger models may be slow or impossible; hardware ages; patching, model management, backup, and endpoint security remain your responsibility.
2 Own GPU server on-premise	Company server room with one or several GPUs; central local inference or RAG service	Sensitive data can remain inside the company network; more capacity than a single workstation; central governance for several users and applications.	High upfront cost; power, cooling, redundancy, maintenance, and specialist skills are needed; capacity may sit idle or be too small during peaks.
3 Own hardware in a colocation centre	Company-owned GPU server or rack hosted in an EU data centre	You still control the hardware and software, while professional facilities provide power, cooling, physical security, and connectivity; easier to operate larger systems than in an office.	You still carry hardware investment and lifecycle risk; remote administration and data-centre fees add complexity; scaling is slower than renting cloud capacity.
4 EU-headquartered cloud / AI provider	Examples: IONOS, OVHcloud, Scaleway, STACKIT	GPU capacity on demand without owning hardware; European contracting and data-residency options can simplify some sovereignty and data-protection questions; easier scaling and managed services.	Compliance still depends on configuration and your use case; service portfolios may be smaller than global hyperscalers; cloud cost can rise quickly for continuous GPU use.
5 Sovereign European variants using global-cloud technology	Examples: AWS European Sovereign Cloud; partner-operated Google Cloud sovereignty offerings such as S3NS or T-Systems	Combines a broad technology ecosystem with stronger EU operational, residency, and governance controls; useful where familiar hyperscaler technology is needed but sovereignty requirements are higher.	Sovereignty varies by service and contractual structure; not every service may be available; the underlying technology and corporate supply chain can still create dependencies outside Europe.
6 Standard EU region / European entity of a US hyperscaler	AWS, Microsoft Azure, or Google Cloud services configured for EU regions or EU data boundaries	Very broad service range, mature tooling, global support, elastic scaling, and rapid access to new AI capabilities; EU data-residency controls are increasingly strong.	EU data location is not the same as full legal or operational sovereignty; some support, telemetry, subprocessors, or exceptional transfers may still require review; lock-in can be significant.
7 Global or US AI cloud service / API	Direct SaaS or frontier-model APIs operated globally or from the US	Fastest access to leading models and features; almost no infrastructure work; excellent for experimentation and highly variable workloads.	Least infrastructure control; contracts, retention, model-training use, international transfers, availability, price changes, and vendor dependence need close attention - especially for sensitive data.

Important: "European", "EU-hosted", "GDPR-compliant", and "sovereign" are not synonyms. Check the specific service, legal entity, operating model, support access, subprocessors, and data flows - not only the location of the data centre.

What Changes as You Move Through the Stack?

Moving from level 1 toward level 7 usually increases elasticity, managed services, and access to the newest models. Moving in the other direction usually increases direct control, offline capability, and independence from an external service. Neither direction is automatically better: the right architecture depends on what the AI is doing and what would happen if data leaked, the provider changed its terms, or the service became unavailable.

One important nuance: even a fully local setup is not completely "European" or independent. GPUs, CPUs, firmware, operating systems, model weights, and software libraries are part of global supply chains. Sovereignty is therefore a gradient - and good architecture makes the dependencies visible rather than pretending they do not exist.

Hybrid Is Often the Practical Answer

For many SMEs, the practical answer is a hybrid architecture. Sensitive documents, embeddings, trade secrets, or a company knowledge base can stay locally or on a controlled European infrastructure, while large cloud models are used only for tasks that do not require the most sensitive information.

A simple routing rule can already help: public or low-sensitivity tasks may use a flexible cloud service; confidential tasks may go to an EU-controlled environment; highly sensitive workloads may stay on-premise. Using standard APIs and portable model formats can also make it easier to change providers later.

Seven Questions to Ask Before Choosing a Provider

1. Where are prompts, uploaded files, logs, embeddings, and outputs stored and processed?
2. Which legal entity signs the contract, and which parent company or non-EU law may still be relevant?
3. Can provider staff or subprocessors access the data, and from which countries?
4. Is customer data used to train or improve shared models, and can this be disabled contractually?
5. Can the model or workload run elsewhere later, or is it tied to proprietary APIs, databases, or orchestration?
6. What are the real costs at steady load: hardware, electricity, administration, GPU hours, tokens, storage, and data transfer?
7. What is the fallback if the model, API, internet connection, or provider is unavailable?

> Regulatory reality check
> Local hosting does not automatically make an AI use GDPR- or AI-Act-compliant, and using a US provider does not automatically make it unlawful. The relevant questions still depend on the data, the purpose, your role, the risk category, the contract, and any international transfer mechanism. For transfers to certified US organisations, the EU-US Data Privacy Framework is one possible transfer mechanism; other safeguards may also be relevant.

> Also worth watching: the EU Data Act
> Vendor lock-in is also a regulatory topic. The EU Data Act already contains cloud-switching and interoperability rules. Switching charges - including relevant data-egress charges - are due to disappear from 12 January 2027. That does not make every workload portable, but it makes exit planning an increasingly reasonable procurement requirement.

Looking Forward

A useful sovereignty strategy does not start with the question "cloud or no cloud?" It starts with a map of data sensitivity, required model capability, uptime, cost, and acceptable dependencies. Once those are clear, SMEs can deliberately choose local, on-premise, European cloud, global cloud - or a combination of them.

Content provided by the Machine Learning Group at RPTU for the Boost AI Monthly Background Newsletter series, September 2026 edition.

Steffen Reithermann <steffen.reithermann@cs.rptu.de> - ml.cs.rptu.de

Newsletter archive: ml.cs.rptu.de/projects/Boost-AI/newsletter/

RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau