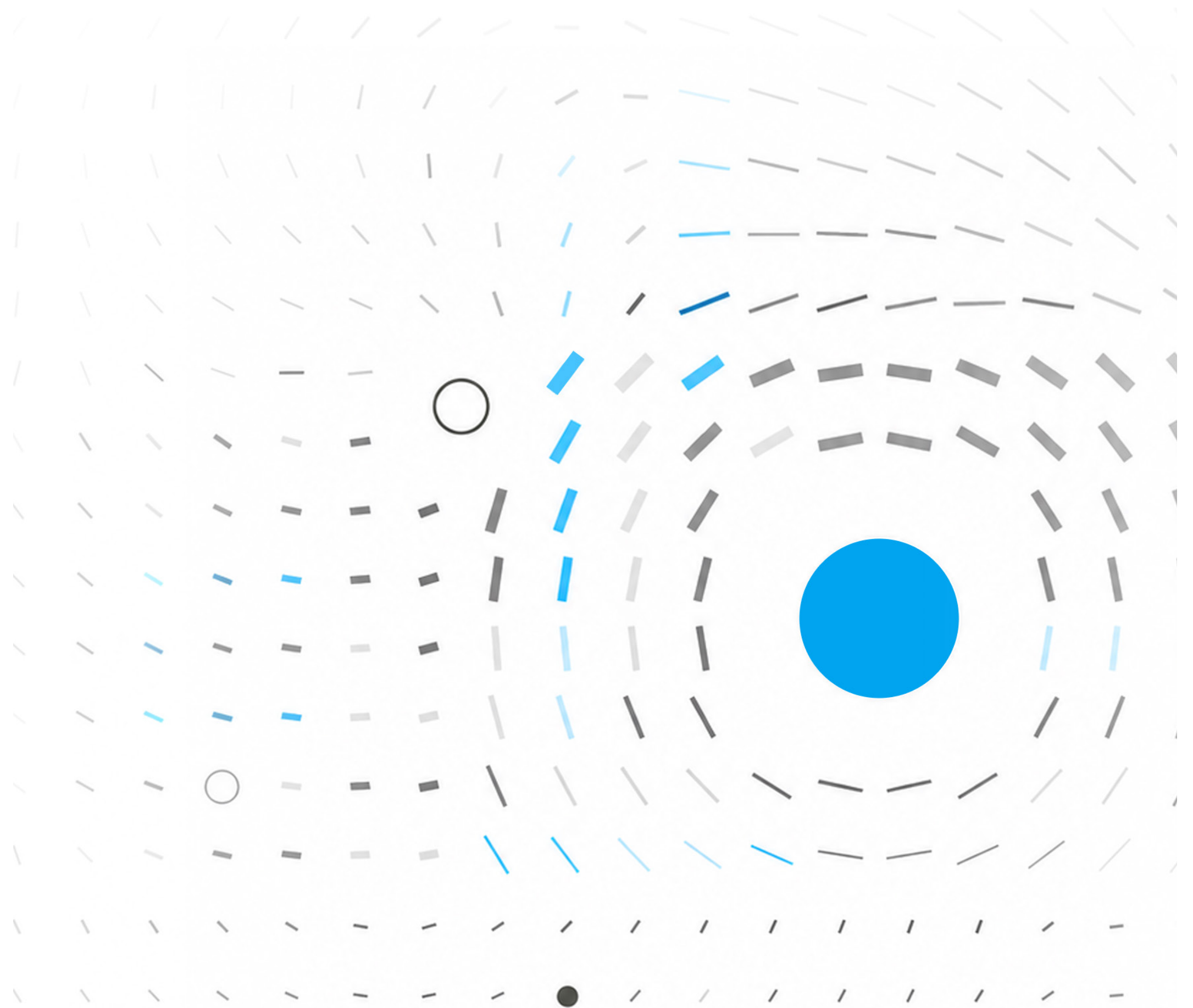


Sovereign AI: Control, choice, flexibility, resilience

How Microsoft helps governments and regulated industries
adopt advanced AI with NVIDIA technologies



Who this paper is for:

Leaders accountable for AI adoption and AI risk, including CIOs, CISOs, CDOs, privacy, compliance and risk leaders, digital and transformation executives, and the policy stakeholders who advise them.

Contents

01

Introduction

02

Defining
sovereign AI

03

Why sovereignty
matters: The forces
driving sovereign AI

04

Control, choice,
flexibility,
and resilience

05

Sovereignty across
the AI stack

06

Sovereign AI
in customer-
controlled
environments

07

Choosing the right
operating model

08

Get started



Introduction

Across Europe and around the world, governments and regulated industries are increasingly using AI to transform public services, strengthen critical infrastructure, improve healthcare and financial systems, advance scientific discovery, and support industrial innovation. As AI becomes embedded in essential operations, organizations need confidence in how sensitive data, systems, and digital capabilities are governed.

AI expands the sovereignty conversation beyond where data is stored. Prompts, retrieval systems, models, agents, outputs, telemetry, and operational tools all become part of an AI workload. As a result, organizations must consider not only data location, but also how AI systems are operated, governed, and controlled.

Sovereignty is multi-dimensional. It includes focus on strengthening local economic engagement, resilience, risk management, and other elements. Organizations must determine what matters most for each workload and apply controls proportionate to their data sensitivity, mission requirements, operational needs, and threat profile.

Key questions for sovereign AI

- Where AI processing occurs
- Who can access and administer the environment
- How data and models are governed
- Who controls encryption and access
- Which external dependencies remain
- Whether controls can be verified
- Whether critical workloads can continue operating when connectivity or operating conditions change

These questions ultimately come down to four principles: control, choice, flexibility, and resilience.

This paper helps government and regulated-industry leaders translate sovereign AI requirements into practical decisions. It outlines the key principles and questions to consider across the AI stack and provides guidance for choosing the right operating model.

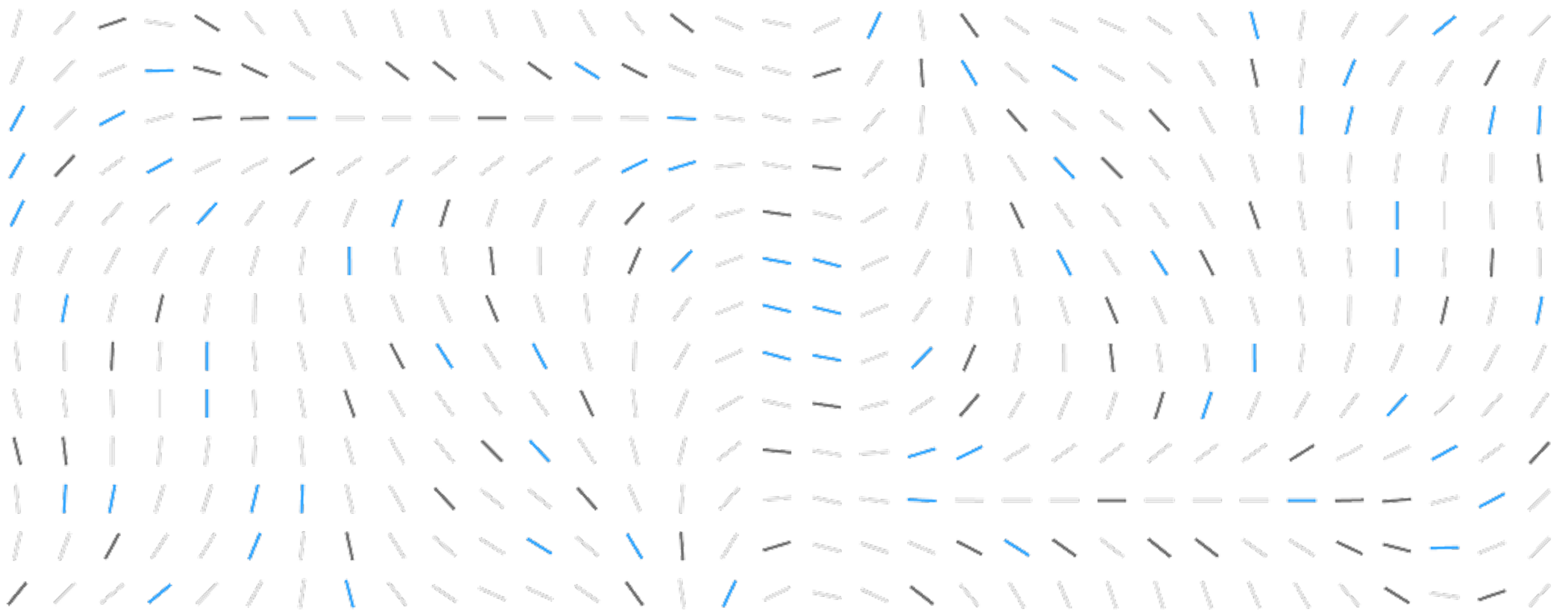
Microsoft definition of sovereign AI

Sovereign AI is the design, deployment, and operation of AI workloads under defined controls for data, access, governance, infrastructure, and operations.

It enables organizations to adopt AI while aligning with the regulatory, operational, security, resilience, and policy requirements that apply to each workload. Those requirements vary by jurisdiction, sector, mission, data sensitivity, and risk profile. Some workloads can meet sovereignty requirements in public cloud environments with appropriate controls, while others require infrastructure and operations to remain within customer-controlled environments.

Microsoft's vision is that organizations should be able to pursue AI innovation while maintaining the control required for sensitive and mission-critical workloads. Organizations can therefore apply different levels of control and operational autonomy based on workload requirements rather than adopting a single sovereignty model across their estate.

Why sovereign AI matters



As AI becomes embedded in essential services and operations, organizations face three connected priorities: demonstrating control, maintaining resilience, and preserving strategic choice.

Control and accountability

Organizations need visibility into where data is processed, where workloads run, who can access and administer them, how models are governed, and how controls are applied across the AI lifecycle. This includes source data, prompts, retrieval systems, embeddings, models, outputs, logs, inference environments, and supporting infrastructure. Clear evidence supports regulatory oversight, audit, internal accountability, and informed risk decisions.

Operational resilience

AI may support services that must continue during connectivity loss, infrastructure failure, cyber incidents, supply-chain disruption, or changes in operating conditions. Organizations therefore need to understand external

dependencies, determine the required level of operational autonomy, and select an environment that supports continuity for each workload.

Innovation and strategic choice

Organizations need continued access to advances in models, accelerated computing, AI software, infrastructure, and partner ecosystems. Model choice, deployment choice, interoperability, and portability help organizations adopt new capabilities while retaining control over where workloads run, how data is managed, and which dependencies they accept.

Principles of Sovereign AI

Microsoft’s approach is organized around four principles. Together, they provide a practical framework for deciding what must remain under an organization’s authority, which technologies best serve the workload, where the workload should operate, and how critical AI capabilities can continue under changing conditions.

Principle	What it means
Control	Retain authority over data, models, infrastructure, operations, and access, with evidence that required safeguards are in place.
Choice	Select the models, platforms, architectures, and operating models that best fit workload requirements.
Flexibility	Deploy AI across public cloud, hybrid, customer-controlled, partner-operated, and disconnected environments as requirements evolve.
Resilience	Maintain critical AI capabilities through changes in connectivity, infrastructure, regulation, and operating conditions.

Together, these principles provide a framework for evaluating sovereign AI requirements and selecting the controls, technologies, and operating models appropriate for each workload.

Control across the AI stack

Organizations need authority over data, models, infrastructure, operations, and access to demonstrate that governance requirements are being met. Microsoft helps organizations establish consistent governance, security, encryption, monitoring, and operational safeguards across sovereign environments. Microsoft’s identity, governance, security, and operational capabilities together with NVIDIA accelerated infrastructure, networking, AI software, and confidential computing help organizations implement and demonstrate control based on each workload’s requirements.

Choice across models and technologies

Different AI workloads require different models, infrastructure, performance characteristics, and operating environments. Microsoft and NVIDIA support a broad model ecosystem while maintaining consistent governance and operational practices. Microsoft Foundry and Foundry Local on Azure Local support model development and deployment across cloud and customer-controlled environments. That ecosystem includes NVIDIA Nemotron models in the Foundry Local on Azure Local catalog, providing organizations with NVIDIA-developed options for reasoning, coding,

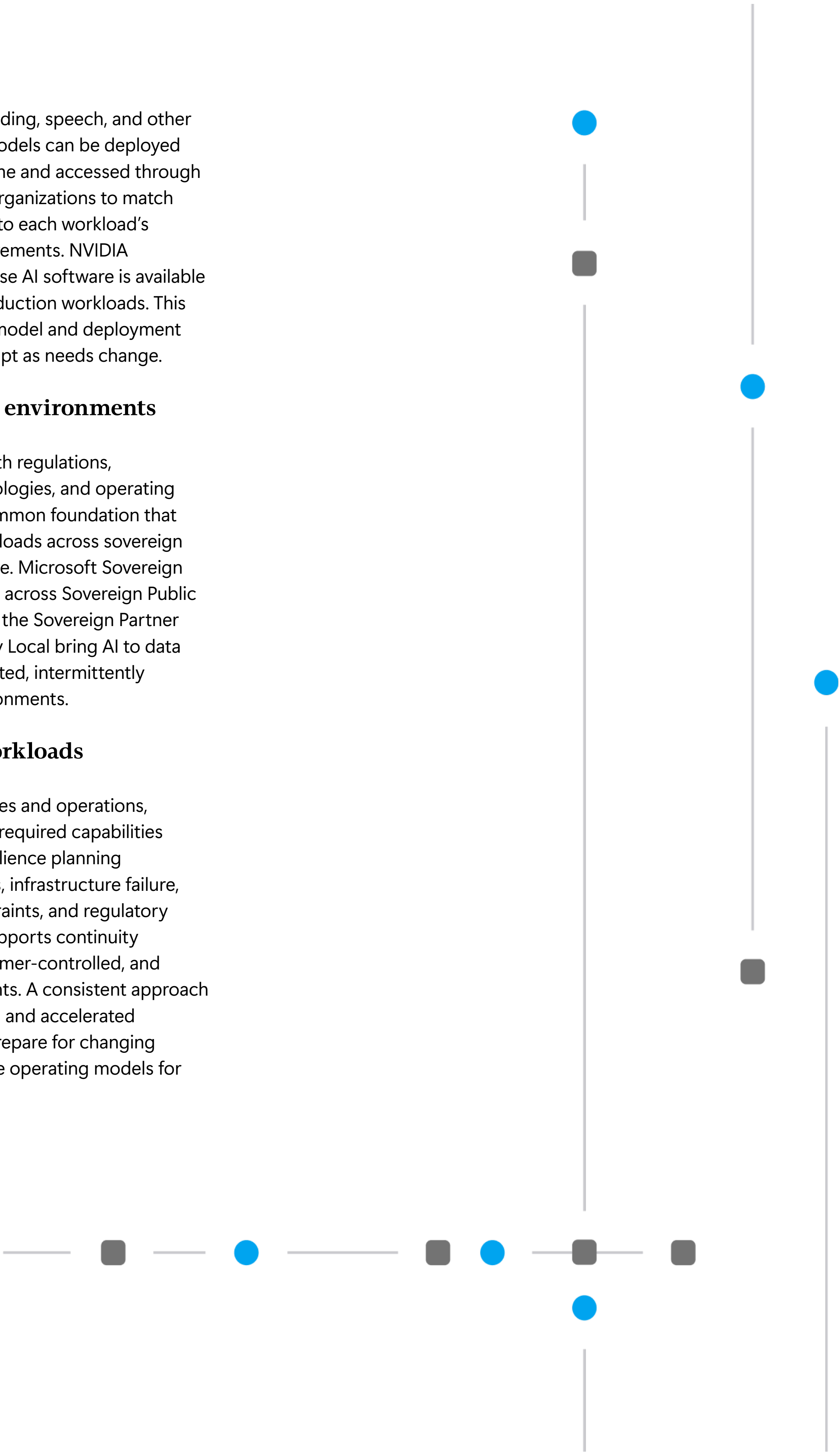
mathematics, multimodal understanding, speech, and other specialized workloads. Nemotron models can be deployed with the GPU-optimized vLLM runtime and accessed through OpenAI-compatible APIs, allowing organizations to match model capability and infrastructure to each workload’s sovereignty and performance requirements. NVIDIA accelerated computing and enterprise AI software is available across the Azure ecosystem for production workloads. This allows organizations to choose the model and deployment environment independently and adapt as needs change.

Flexibility across sovereign environments

Sovereignty requirements evolve with regulations, missions, data classifications, technologies, and operating conditions. Microsoft provides a common foundation that enables organizations to move workloads across sovereign environments as requirements evolve. Microsoft Sovereign Cloud provides deployment choices across Sovereign Public Cloud, Sovereign Private Cloud, and the Sovereign Partner Ecosystem. Azure Local and Foundry Local bring AI to data that cannot move, including connected, intermittently connected, and disconnected environments.

Resilience for critical AI workloads

As AI becomes part of critical services and operations, organizations need confidence that required capabilities can continue during disruption. Resilience planning should account for connectivity loss, infrastructure failure, cyber incidents, supply-chain constraints, and regulatory or operational change. Microsoft supports continuity planning across cloud, hybrid, customer-controlled, and disconnected operating environments. A consistent approach to governance, security, AI software, and accelerated infrastructure helps organizations prepare for changing conditions without creating separate operating models for every environment.



Sovereignty across the AI Stack

Sovereign AI spans multiple layers, from the applications users interact with to the infrastructure that powers them. Organizations should evaluate control, governance, resilience, and risk across the entire stack rather than focusing on any single component in isolation.

The sovereignty questions organizations ask about AI often map to different layers of the stack. Understanding these layers helps identify where controls, governance, resilience, and operational decisions should be applied.

AI Layer	Sovereignty Consideration
Applications and experiences	Start with the business outcome. Understand what data an AI application can access, what actions it can take, and how those interactions are governed.
AI and model platform	Determine which models may be used, where they run, how they are evaluated and updated, and how access is governed. Microsoft Foundry, Foundry Local, and NVIDIA AI Enterprise software support development and production deployment across cloud and customer-controlled environments.
Data and application platform	Decide where data, applications, and AI processing should operate based on jurisdiction, latency, resilience, and connectivity. Azure and Azure Local provide platforms across cloud and customer-controlled environments.
Security, governance, and sovereignty controls	Decide where data, applications, and AI processing should operate based on jurisdiction, latency, resilience, and connectivity. Azure and Azure Local provide platforms across cloud and customer-controlled environments.
Accelerated infrastructure	Match computing and networking capacity to workload performance and control requirements

This stack-wide view establishes the foundation for the next section, which examines how Microsoft Sovereign Private Cloud supports AI workloads that must remain within customer-controlled environments.

Sovereign AI in customer-controlled environments

For some AI workloads, the sovereignty requirement is clear: data, infrastructure, and operations must remain within an environment the organization controls. This is common for mission-critical government systems, clinical environments, industrial operations, critical infrastructure, and highly regulated financial workloads. These workloads often represent the most demanding end of the sovereign AI continuum.

Built on Azure Local and accelerated by NVIDIA, Microsoft Sovereign Private Cloud brings cloud, AI, security, and data services into customer-controlled environments. With Foundry Local on Azure Local, organizations can deploy selected NVIDIA Nemotron models on GPU-enabled, Azure Arc-managed Kubernetes infrastructure close to sensitive data. Model artifacts are cached and served locally, helping keep inference processing within the organization's controlled operational boundary. Together, these capabilities enable organizations to run advanced AI close to sensitive data while maintaining the control, resilience, and operational independence these workloads require.

What Sovereign Private Cloud enables:

- **Operate anywhere:** Run AI in connected, intermittently connected, or fully disconnected environments and adapt as requirements change. .
- **Scale as needed:** Deploy from small, edge environments to large-scale AI infrastructure, with support for continuity and disaster recovery.
- **Preserve infrastructure choice:** Use validated hardware platforms and NVIDIA GPU-enabled systems from a broad ecosystem of providers
- **Bring AI to the data:** Run data, AI, and applications within the same operational boundary, minimizing the need to move sensitive information.
- **Manage as one estate:** Apply consistent governance, security, and operations across distributed environments through Azure Arc.

Implementation guidance and recommendations

Organizations evaluating sovereign AI should begin with the workload, not the technology. A public-facing assistant, employee agent, clinical AI system, industrial application, or mission-critical government service may use similar AI technologies but require very different approaches to control, resilience, and governance.

A practical starting point is to evaluate each workload across five dimensions:

Consideration	Key Questions
Data and processing boundaries	What data does the workload use, and where can it be processed and stored?
Models and AI governance	Which models are approved and how will they be governed, monitored, and updated?
Security and access	What controls are required for identities, encryption, infrastructure, and administrative access?
Deployment and operations	Which deployment model best meets requirements for control, resilience, connectivity, and performance?
Evidence and accountability	What monitoring, audit, and reporting capabilities are needed to demonstrate compliance and oversight?

Match the workload to an operating model

Most organizations will require multiple operating models across different workloads. The appropriate choice depends on workload requirements.

- **Sovereign Public Cloud** is often appropriate when organizations require cloud-scale innovation, broad AI capabilities, and sovereignty controls that can be met through cloud-native governance, security, and operational safeguards.
- **Sovereign Private Cloud** is appropriate when infrastructure, operations, or AI processing must remain within customer-controlled environments, including connected or disconnected deployments.
- **Sovereign Partner Ecosystem** can support country-specific requirements that call for local operation, national oversight, or specialized sovereignty expertise.

Build for change

Sovereignty requirements, regulations, and AI technologies will continue to evolve. Organizations benefit from a consistent architecture that allows workloads to move across sovereign environments without redesigning applications or adopting separate operating models.

Microsoft with NVIDIA technologies help provide that flexibility through a common approach to identity, security, governance, AI development, and accelerated infrastructure across the sovereign environments. Organizations can adapt deployment models as requirements change while maintaining consistent governance, security, and operational practices.

Getting started

Organizations should start with a small set of high-value AI workloads and evaluate them through a consistent framework for control, governance, resilience, and risk. As requirements and confidence evolve, they can expand adoption across additional workloads and operating environments.

Microsoft believes organizations should not have to choose between AI innovation and sovereignty. By combining control, choice, flexibility, and resilience, organizations can adopt advanced AI while maintaining confidence in how data, models, infrastructure, and operations are governed.

Learn more

aka.ms/sovereign

aka.ms/azurelocal

nvidia.com

Explore Sovereign Private Cloud:

aka.ms/sovereignprivatecloud

[https://www.nvidia.com/en-us/data-center/
gpu-cloud-computing/microsoft-azure/](https://www.nvidia.com/en-us/data-center/gpu-cloud-computing/microsoft-azure/)

