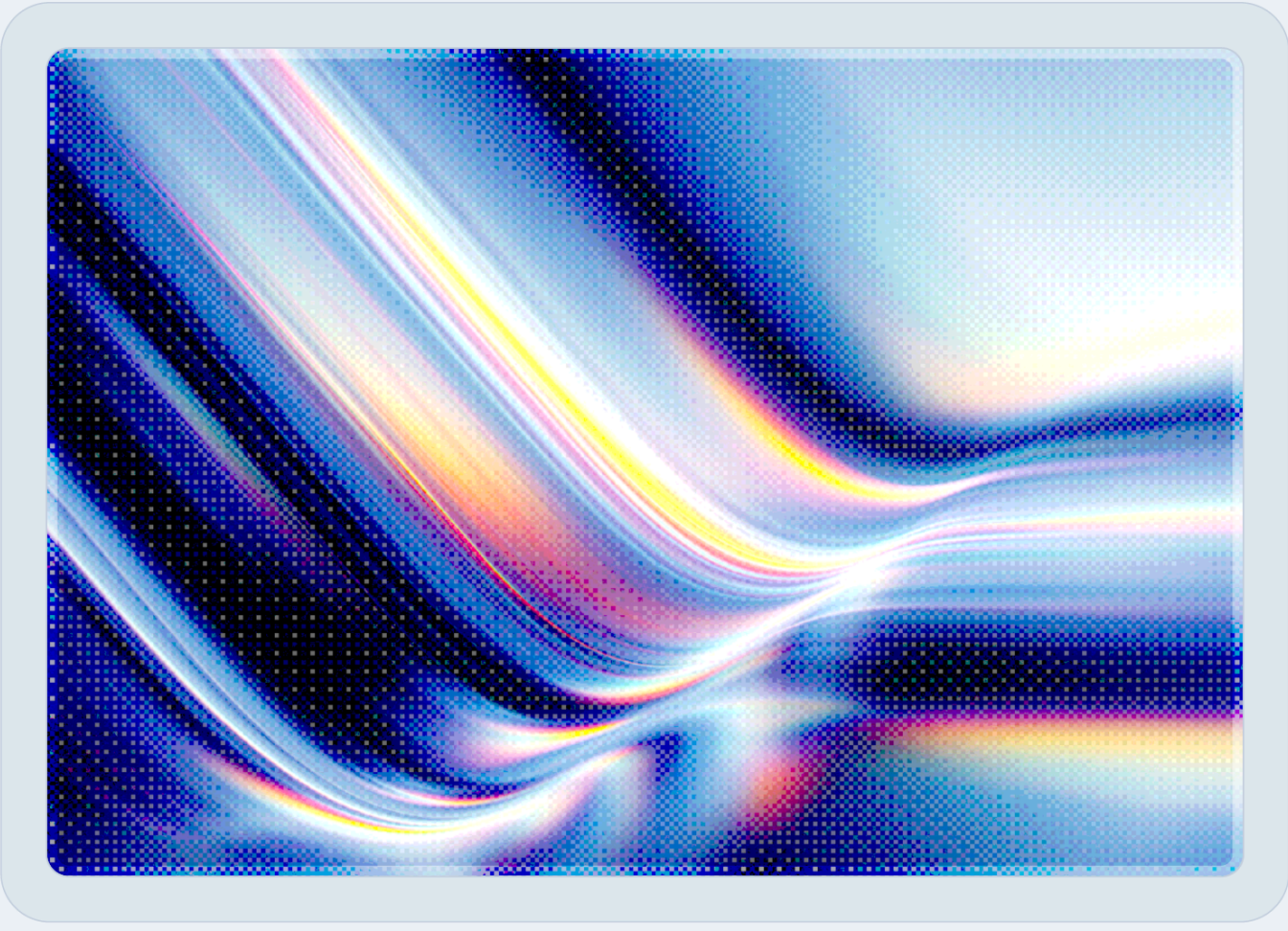
CSAI Foundation | Cloud Security Alliance

Sovereign AI Access Controls and Frontier Model Dependency Risk

Managing Enterprise Concentration Risk in the Age of Foundation Model APIs

2026-06-27

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Table of Contents

Executive Summary 4

Section 1: The Frontier Model Concentration Landscape 6

- The Dependency Curve
- Provider Reliability and the SLA Gap

Section 2: Sovereign AI — Definitions, National Strategies, and the Enterprise Dimension 8

- What Sovereign AI Means
- National Sovereign AI Initiatives
- The Enterprise Sovereignty Imperative

Section 3: Access Control Failure Modes in Frontier AI Deployments 11

- Data Exfiltration via Prompt Construction
- Authentication, Authorization, and the Credential Management Gap
- Shadow AI and Governance Visibility

Section 4: The Regulatory Environment 14

- EU AI Act — Obligations Already in Force
- GDPR Enforcement Actions Against AI API Usage
- Emerging EU Regulatory Frameworks for AI Data Sovereignty
- U.S. Export Controls on AI Model Weights and Compute

Section 5: A Framework for Sovereign AI Access Controls 17

- The AI Gateway as Enterprise Control Plane
- Identity and Access Management for AI Systems
- Data Residency Architecture
- Multi-Provider Resilience Architecture
- Zero Data Retention and Encryption Controls

Section 6: CSA Resource Alignment 20

Section 7: Conclusions and Recommendations 22

- Immediate Actions (0–90 Days)
- Near-Term Mitigations (90 Days to One Year)
- Strategic Considerations

References 24

Executive Summary

The enterprise AI stack has quietly reorganized itself around a small number of frontier model providers. According to Menlo Ventures' year-end 2025 survey of enterprise technology spending, three providers — Anthropic, OpenAI, and Google — collectively hold approximately 88 percent of enterprise large language model API spending, with Anthropic claiming 40 percent, OpenAI 27 percent, and Google 21 percent [1][2]. Enterprise generative AI spending tripled in a single year, reaching $37 billion in 2025 [2]. By the time those figures were published, both Ramp's May 2026 business spend index and Gartner's August 2025 forecast had confirmed the trajectory: Anthropic and OpenAI together reached the majority of U.S. businesses paying for AI services [3], while Gartner predicted that 40 percent of enterprise applications would embed task-specific AI agents by end of 2026, up from under 5 percent in 2025 [18].

That concentration creates a category of structural risk that traditional third-party risk management frameworks were not designed to assess [20]. When a critical database vendor has an outage, the blast radius is typically bounded by the set of applications that depend on that database. When a frontier AI model provider has an outage — as OpenAI did on June 10–11, 2025, in a 34-hour disruption affecting every service it operates, or as Anthropic did in June 2026 when a sub-agent multiplication bug cascaded across 8,000 or more reported incidents — the blast radius is every enterprise workflow that has been rebuilt around that provider's API [7][8]. A Zapier survey found that 47 percent of enterprise leaders believe at least one key business function would stop working if their primary AI vendor experienced significant downtime or a major policy change; only 6 percent say they could switch AI vendors without material disruption [4]. IBM research has similarly found that 91 percent of executives do not fully understand their organization's AI dependencies across vendors and infrastructure, and 71 percent say switching their primary AI vendor would be difficult [5]. Neither OpenAI nor Anthropic has historically met the 99.9 percent uptime threshold standard in enterprise software SLAs across their full service surface [33].

Alongside availability risk sits a set of access control and data governance challenges that are structurally different from those arising with conventional cloud services. When an enterprise sends prompts to a frontier AI API, it transmits full prompt content — which routinely includes proprietary business context, personally identifiable information, and sometimes regulated data — to infrastructure it does not control, under terms that many security and compliance teams have not reviewed carefully. The Samsung case, in which three separate employees leaked proprietary semiconductor source code and internal meeting transcripts to ChatGPT within a 20-day period before Samsung detected and banned generative AI tools company-wide, is illustrative but far from isolated [12]. Italy's data protection authority fined OpenAI €15 million in November 2024 for GDPR violations including failure to establish a lawful basis for training data processing and insufficient transparency obligations — a decision that was annulled in March 2026 on jurisdictional grounds but whose substantive privacy findings remain unresolved pending a forthcoming Irish

Data Protection Commission investigation [9][37]. The European Data Protection Board's December 2024 Opinion 28/2024 on AI models added a further concern: if an AI model was trained on unlawfully processed personal data, that taint may extend downstream to enterprises that deploy the model via API, potentially implicating the integrating organization in the underlying violation [11].

Sovereign AI — the set of national and enterprise-level initiatives designed to ensure that AI capabilities remain within jurisdictional control and under domestically governed access policies — is gaining both policy urgency and commercial traction. Gartner forecasts that more than one-third of enterprises will use localized AI platforms by 2027, up from roughly 5 percent today [28]. National governments have announced sovereign AI commitments spanning French infrastructure investments of €109 billion under the France 2030 plan, South Korea's deployment of 260,000 GPUs across domestic sovereign clouds, and the European Commission's funding of five AI gigafactories under the EU Chips Act [19]. The United States Bureau of Industry and Security meanwhile established export controls on AI model weights under the Export Administration Regulations in January 2025 — a policy that the subsequent administration paused in May 2025 pending replacement, but whose underlying logic of treating frontier model weights as strategic assets subject to access controls has persisted into revised 2026 guidance [15][16][32].

This paper argues that enterprises cannot fully delegate access control to their AI providers, that the regulatory and operational environment now requires a systematic enterprise-side governance layer, and that CSA's existing frameworks — the AI Controls Matrix, the Agentic Trust Framework, MAESTRO, and Zero Trust guidance — provide the architecture for building that layer. The paper closes with a prioritized set of immediate actions, near-term mitigations, and strategic recommendations calibrated to the pace at which enterprise AI deployments are deepening.

Section 1: The Frontier Model Concentration Landscape

The Dependency Curve

Enterprise dependence on frontier AI models is accelerating faster than governance frameworks designed to manage that dependence. The Menlo Ventures enterprise AI market survey, published in December 2025, measured enterprise large language model API spending at $8.4 billion in the middle of that year and projected full-year enterprise generative AI spending at $37 billion — a figure 3.2 times larger than 2024's $11.5 billion [1][2]. The sources underlying that growth are structurally concentrated: three providers account for 88 percent of enterprise LLM API spending, and no fourth-place competitor holds more than a few percent of the remainder [2]. Ramp's May 2026 index, drawing on transactional data from more than 50,000 U.S. companies, confirmed that Anthropic and OpenAI had achieved roughly equal business adoption rates at 34.4 and 32.3 percent respectively, while Google's business AI penetration remained below 5 percent for most of the prior year [3][26]. Deloitte's 2026 State of AI survey found that 74 percent of all AI-generated economic value flows to just 20 percent of organizations, with the remaining enterprises facing a compounding disadvantage as agentic AI capabilities accelerate [36].

Diversification is occurring at the testing and evaluation layer, but not at the production-workflow layer. MindStudio research from 2026 found that 81 percent of enterprises now evaluate or run three or more model families, and that 79 percent of OpenAI users also pay for Anthropic products [35]. But multi-provider evaluation does not translate into multi-provider resilience: most enterprises continue to route their most critical and highest-volume workflows through a single primary provider, concentrating operational dependence even as nominal portfolio diversification grows. The abstraction layers that would enable rapid failover — model-agnostic API wrappers, provider-agnostic prompt engineering, and maintained secondary-provider credentials — are present in fewer than 20 percent of production enterprise AI deployments, according to industry estimates [35].

Agentic AI is amplifying the concentration risk. As enterprises move beyond discrete inference calls toward multi-step agentic workflows that chain model calls together, the dependency on any single provider's API reliability deepens. A workflow in which a model performs reasoning, tool invocation, code generation, and output synthesis across ten sequential calls is ten times as sensitive to provider downtime as a single inference request. Gartner projects that 40 percent of enterprise applications will embed task-specific AI agents by end of 2026, up from under 5 percent in 2025 [18]. Gartner simultaneously projects that more than 40 percent of agentic AI projects will be canceled by end of 2027, citing escalating costs, unclear business value, and inadequate risk controls — a pattern that suggests enterprises are building agentic dependencies faster than they are building the governance structures to manage them [17].

Provider Reliability and the SLA Gap

Frontier AI model providers operate infrastructure at a reliability level that is materially below the 99.9 percent uptime threshold standard in enterprise software service level agreements. An independent analysis of OpenAI's operational history found baseline uptime hovering around 99.3 percent, implying approximately five hours of downtime per month — a number that understates the impact when the downtime coincides with peak business hours or affects multiple services simultaneously [33]. Anthropic's 90-day status metrics as of mid-2026 reported claude.ai at 99.12 percent, Claude Code at 99.28 percent, and the API at 99.41 percent, all below the standard enterprise SLA threshold [8].

The most significant documented disruptions occurred in rapid succession. On June 10–11, 2025, OpenAI suffered an extended global outage lasting 34 hours that affected ChatGPT, Sora, the Agents API, the Realtime Speech API, and all related developer services simultaneously [7]. The incident followed a cluster of other 2024 disruptions, including a nine-hour outage in December attributable to Microsoft Azure infrastructure and a separate four-plus-hour outage on December 11, 2024, caused by a configuration error that drove error rates across all services above 90 percent [33]. Anthropic's June 2026 disruption had a different technical character — a bug in Claude Code's sub-agent architecture triggered exponential sub-agent multiplication, cascading failures across the developer console and API — but the business consequence for enterprises that had built automated pipelines atop the API was equivalent: unexpected, uncontrolled, and difficult to recover from without a failover capability that most deployments did not have [8].

The gap between provider reliability and enterprise SLA expectations is not primarily a criticism of provider engineering. It reflects a structural mismatch: frontier model providers are operating novel infrastructure categories at scales and complexities that outpace the reliability engineering practices developed for conventional web services, while enterprises are adopting their APIs at a pace that treats them as if they were utility-grade cloud primitives. Resolving the gap requires both continued provider investment in reliability engineering and enterprise investment in multi-provider architecture, circuit breakers, graceful degradation, and the access control infrastructure needed to manage credentials and traffic across multiple providers simultaneously.

Section 2: Sovereign AI – Definitions, National Strategies, and the Enterprise Dimension

What Sovereign AI Means

"Sovereign AI" is used to describe at least three distinct but related concepts, and the ambiguity is worth resolving before building governance recommendations around it. At the national level, sovereign AI refers to a country's capacity to develop, deploy, and regulate AI systems without strategic dependence on foreign technology providers – encompassing compute infrastructure, training data, model development, and the regulatory frameworks that govern all of the above [24][25]. At the enterprise level, sovereign AI describes an organization's ability to deploy AI capabilities under its own access policies, within its own governance framework, with sufficient control over data flows, model behavior, and operational continuity to meet its regulatory obligations and risk tolerance. These two definitions overlap but are not identical; an enterprise can achieve meaningful operational sovereignty by deploying open-weight models on cloud infrastructure it controls, even if that infrastructure relies on hardware manufactured abroad.

The Stanford Human-Centered AI Institute's analysis of AI sovereignty notes that the definitional ambiguity creates a risk: "sovereignty" can become a label that governments attach to a wide range of policies – data localization, domestic ownership requirements, AI licensing regimes – without a coherent theory of which dependencies actually create strategic vulnerability [24][27]. For enterprises, the practical question is narrower and more tractable: what dependencies on AI model providers create operational, regulatory, or competitive risks that the enterprise cannot accept, and what controls reduce those risks to acceptable levels?

National Sovereign AI Initiatives

The pace of national sovereign AI investment accelerated through 2025 and 2026. McKinsey estimates that 23 new sovereign AI infrastructure projects were announced globally in the final quarter of 2025, with investment heavily concentrated in the Middle East and East Asia [19]. France committed €109 billion to AI infrastructure under the France 2030 plan, targeting deployment of 1.2 million GPUs and training of 100,000 AI professionals annually by 2030 [19]. South Korea announced plans in late 2025 to deploy more than 260,000 GPUs across domestic sovereign clouds and AI factories in partnership with NVIDIA and domestic technology companies [28]. The European Commission's AI Chips Act is funding five AI gigafactories specifically for training large AI models, and a 100-billion-parameter European open-source language model (SOOFI) is targeted for public release in Q3 2026 [28]. Global spending on sovereign AI systems is projected to surpass $100 billion in 2026 [28].

The United States has taken a different regulatory posture, focusing on controlling the diffusion of frontier AI capabilities abroad rather than building a domestic sovereign stack exclusively for national use. The Bureau of Industry and Security published export controls on January 15, 2025, establishing Export Administration Regulations jurisdiction over advanced computing items including, for the first time, model weights for the most capable AI models [15]. The subsequent administration paused enforcement of the Biden-era AI Diffusion Rule in May 2025 pending a replacement framework, while simultaneously issuing new guidance that extends EAR jurisdiction over AI-related transactions in specific circumstances [16]. In January 2026, BIS issued a revised licensing policy for advanced chips destined for China and Macau, moving from a presumption of denial to case-by-case review for chips at approximately the NVIDIA H200 tier and below [32]. The net effect is a regulatory environment in which both the model weights that power frontier AI and the hardware used to run those models are subject to export licensing requirements in circumstances that affect enterprise procurement and deployment decisions.

The Enterprise Sovereignty Imperative

For enterprise security and compliance teams, the sovereign AI debate translates into five concrete operational requirements. First, data residency: regulated industries in most major jurisdictions — financial services, healthcare, government contractors, telecommunications — face explicit requirements limiting where certain categories of data can be processed and stored. Real-time inference calls to frontier AI APIs constitute data processing, and the jurisdiction in which that inference occurs is often not the jurisdiction in which the enterprise's regulated customers or data subjects are located. OpenAI expanded data residency to cover at-rest storage in ten regions as of 2025, but inference processing for API calls continues to default to U.S. infrastructure as of mid-2026 — meaning that an enterprise can achieve data storage residency while still processing inference in a jurisdiction that violates local data localization requirements [30].

Second, access control accountability: when an enterprise deploys an AI capability via a third-party API, the API provider controls the authentication and authorization mechanisms, the rate limiting and abuse prevention systems, and the audit logging infrastructure. The enterprise's ability to implement least-privilege access, attribute-based authorization, and granular audit trails for AI interactions depends on what the provider exposes — and what the provider exposes varies substantially across providers and tiers. Third, data retention and training: the default data retention policies of frontier AI providers range from seven days for Anthropic API calls (reduced from 30 days in September 2025) to 30 days for OpenAI API calls [31]; all three major providers prohibit using API customer data for model training by default, but the controls that enforce that prohibition — including zero-data-retention arrangements that must be separately negotiated via enterprise agreements — are not universally deployed [31]. Fourth, business continuity: as documented in Section 1, the gap between frontier AI provider uptime and enterprise SLA requirements is significant and creates unmitigated business continuity risk for organizations that have concentrated workflows on a single provider. Fifth, supply chain governance: the EDPB's Opinion 28/2024 makes explicit that enterprises

integrating AI models via API may inherit regulatory exposure if the underlying model was trained using unlawfully processed personal data, establishing a form of AI supply chain liability that has no direct analogue in conventional software procurement [11].

Section 3: Access Control Failure Modes in Frontier AI Deployments

Data Exfiltration via Prompt Construction

The most widely documented enterprise access control failure mode in frontier AI deployments is not an attack but an operational behavior: employees and automated processes transmitting sensitive proprietary and personal data to external AI APIs without authorization or awareness of the implications. The Samsung incident — in which three separate semiconductor engineers sent proprietary source code, defect detection algorithms, and internal meeting transcripts to ChatGPT within a 20-day window before the company's internal security team detected and responded to the pattern — illustrates the speed at which data can leave an enterprise's control when AI tools are adopted faster than governance [12]. Samsung's subsequent company-wide ChatGPT ban triggered similar restrictions at Apple, JPMorgan Chase, Verizon, and Amazon, demonstrating that the Samsung pattern was treated by peer organizations as a systemic risk, not an isolated incident [12].

The structural challenge is that prompts sent to frontier AI APIs routinely contain context that employees regard as necessary to get useful responses. Business strategy documents pasted to get summaries, customer records submitted for analysis, proprietary codebases uploaded for refactoring, and clinical notes submitted for documentation are each individually defensible from the perspective of the employee trying to do their job efficiently, and each individually creates a regulatory and competitive exposure that the employing organization has not consented to. Fifty-five percent of enterprise AI inference is now performed on-premises, with data residency compliance as the primary driver — a figure that implies the remaining 45 percent flows to external providers without the same controls, and that the organizations making on-premises investments are doing so precisely because they identified the risks of external inference [28].

Authentication, Authorization, and the Credential Management Gap

Frontier AI model APIs typically authenticate via long-lived API keys rather than the short-lived, scoped tokens that Zero Trust access control principles recommend. An API key embedded in an application codebase or stored in an environment variable has no inherent expiration, no binding to a specific principal identity, and no attribute-based scope that limits which data the key can be used to process. When that key is shared across multiple applications or teams — as is common in enterprise deployments that began with a single API key and expanded over time — the organization loses the ability to attribute specific AI inference activity to specific principals, enforce least-privilege access at the API level, or revoke access granularly if a specific team's key is compromised. The OmniGPT breach of 2025, in which an attacker gained access to

30,000 user email addresses, phone numbers, and 34 million lines of chat messages including embedded API keys and credentials, illustrates how AI platform breaches can cascade into credential exposure that extends far beyond the immediate platform [28].

The authorization dimension is equally underdeveloped. Conventional access control frameworks limit what data a principal can read or write within a system boundary. When an AI model API is involved, a principal can effectively write a query that causes the AI to synthesize, reason about, or expose patterns across data that the principal is not directly authorized to see — particularly in retrieval-augmented generation configurations where the model has access to a broad document corpus. CSA's guidance on shadow access risks identifies this pattern explicitly: AI model interactions that bypass conventional authorization controls represent a new category of shadow access that sits outside the visibility of traditional identity and access management systems [34]. Enterprises that have deployed frontier AI APIs without extending their identity and access management posture to cover those APIs have effectively created a shadow access boundary that their conventional security controls cannot see.

Survey data from 2025 and 2026 documents the scale of the gap. The 42Crunch State of API Security 2026 report, drawing on real-world production vulnerabilities observed across 2024 and 2025, found that 57 percent of AI-powered APIs were externally accessible, and 89 percent used insecure authentication methods such as static keys [39]. A VentureBeat survey of enterprise security teams found that 88 percent of organizations had confirmed or suspected AI agent security incidents in the preceding year, yet only 22 percent of teams treat AI agents as independent, identity-bearing entities with their own provisioned credentials, and 45.6 percent still rely on shared API keys for agent-to-agent authentication [40]. The credential management problem extends to software development workflows: GitGuardian's March 2026 State of Secrets Sprawl report found that AI service credentials exposed in public GitHub repositories surged 81 percent year-over-year in 2025, reaching more than 1.2 million exposures — and that commits made with AI coding assistance leaked secrets at approximately 3.2 percent, twice the baseline rate for manually authored commits [41].

Shadow AI and Governance Visibility

The governance visibility problem extends beyond access control to include the fundamental question of which AI tools employees are using. Enterprise AI governance programs typically attempt to manage an approved set of tools and APIs, but employees routinely adopt AI capabilities that fall outside that approved set. Skyhigh Security's enterprise security predictions for 2026 identify this as a primary concern: employees are adopting AI tools faster than governance can keep up, and without strong data protections, organizations lose visibility into what sensitive information is being shared, where it goes, and how it persists [28]. Aon's research found that 88 percent of organizations use AI in at least one business function, but only 8 percent have a comprehensive AI governance framework covering that usage [6].

The governance gap has a specific technical manifestation: because frontier AI APIs are accessed over the internet via standard HTTPS, conventional network security controls — firewalls, proxy servers, and DLP systems configured for known data destinations — may not detect or log the full content of AI API calls. Organizations that rely on network-layer visibility for data loss prevention have a structural blind spot for AI-mediated data egress that requires explicit remediation.

Section 4: The Regulatory Environment

EU AI Act — Obligations Already in Force

The EU AI Act establishes obligations for AI providers and deployers that are now partially in effect and will reach full force in 2026. Prohibitions on specific AI practices took effect in February 2025. General-purpose AI model obligations, which apply to frontier model providers serving EU markets, took effect August 2, 2025, requiring providers to maintain technical documentation, publish training data summaries, implement copyright compliance policies, and share information with downstream deployers [13][14]. The full set of high-risk AI system obligations applies from August 2, 2026, with fines reaching €35 million or 7 percent of global annual revenue for the most serious violations [14].

The EU AI Act's extraterritorial scope captures U.S. enterprises and their AI API providers in circumstances that enterprise legal and compliance teams need to evaluate carefully. Article 2(1)(c) applies the Act to any provider or deployer established in a third country where the AI system's outputs are used in the EU, without requiring EU incorporation, physical presence, or EU-hosted servers [14]. An enterprise based in the United States whose AI-assisted customer service system serves EU customers, or whose HR AI tool is deployed for EU-based employees, falls within the Act's scope for those use cases. The downstream API integrator's exposure under the Act includes not just the deployer's own obligations but, per EDPB Opinion 28/2024, potential liability inherited from the provider's training data practices [11].

GDPR Enforcement Actions Against AI API Usage

The GDPR enforcement record on generative AI as of mid-2026 is shorter than many enterprises anticipated, but the direction of travel is clear. Italy's data protection authority issued a €15 million fine against OpenAI in November 2024 for GDPR violations including processing personal data without a lawful basis from ChatGPT's November 2022 launch through March 2023, insufficient transparency, failure to notify the regulator of a data breach, and inadequate age verification [9][21]. The Rome court annulled the decision in March 2026 on jurisdictional grounds — OpenAI's February 2024 establishment of OpenAI Ireland Limited made the Irish Data Protection Commission the lead supervisory authority under GDPR's one-stop-shop mechanism — but the substantive privacy findings that drove the fine remain substantively uncontested and are now the subject of ongoing Irish DPC investigation [37].

The European Data Protection Board's May 2024 interim report from its ChatGPT Taskforce acknowledged that OpenAI relies on legitimate interests under Article 6(1)(f) as its primary legal basis for training data processing, and that the three-part necessity, purpose, and balancing test required to validate that basis had not yet been assessed by a lead authority [10]. EDPB Opinion 28/2024, adopted in December 2024,

added the downstream liability concern: an AI model trained on unlawfully processed personal data may itself be unlawful to deploy, and enterprises that integrate unlawfully trained models via API may inherit that exposure [11]. The EU-U.S. Data Privacy Framework, which provides the primary transfer mechanism for EU personal data flowing to U.S.-based AI infrastructure, survived a legal challenge before the EU General Court in September 2025, but a pending Court of Justice of the European Union appeal (Case C-703/25 P) and the ongoing uncertainty around the U.S. Privacy and Civil Liberties Oversight Board's institutional status mean that DPF adequacy cannot be treated as settled [37].

Emerging EU Regulatory Frameworks for AI Data Sovereignty

Beyond the AI Act, two additional regulatory developments in the European Union are reshaping the landscape for enterprises using U.S.-based frontier AI APIs. First, the European Commission's proposed Cloud and AI Development Act (CADA) would establish four tiered sovereignty assurance levels for cloud and AI services used by EU public sector bodies, ranging from Level 1 (EU data storage only) through Level 4 (full EU software supply chain control). Under the proposed framework, U.S.-based AI API providers including OpenAI, Anthropic, and Google cannot qualify above Level 1, because their infrastructure remains subject to U.S. law — specifically the CLOUD Act, which authorizes U.S. law enforcement to compel disclosure of data stored by U.S.-incorporated entities regardless of geographic location of the servers. The Commission demonstrated the CADA framework's practical application in April 2026, awarding a €180 million sovereign cloud contract exclusively to European providers [42]. Second, the EU-U.S. Data Privacy Framework, while surviving its first judicial challenge in September 2025, faces an ongoing appeal before the Court of Justice of the European Union; enterprises should maintain Standard Contractual Clauses as a fallback transfer mechanism for AI API data flows, given the DPF's uncertain long-term status [37].

The CLOUD Act dimension deserves explicit attention in enterprise AI risk assessments. When an enterprise contracts with a U.S.-incorporated AI API provider — even if that provider stores and processes data in EU-located infrastructure under a data residency agreement — U.S. law enforcement may compel the provider to produce that data without notifying the enterprise or the relevant EU data protection authority. This structural tension between EU data sovereignty requirements and U.S. jurisdictional reach over U.S.-incorporated entities cannot be resolved through data residency agreements alone; it requires either contracting with EU-incorporated AI providers, using open-weight models on EU-controlled infrastructure, or obtaining explicit legal advice on the CLOUD Act risk for the specific data categories at issue.

U.S. Export Controls on AI Model Weights and Compute

The January 2025 BIS export control rule established, for the first time, EAR jurisdiction over the model weights of the most capable AI models [15]. While the Biden administration's AI Diffusion Rule framework was rescinded by the Trump administration in May 2025, BIS issued concurrent guidance making clear that AI-related transactions continue to face licensing review in specific circumstances, particularly where they

involve jurisdictions subject to enhanced scrutiny [16]. The January 2026 BIS revision to export licensing policy for advanced chips destined for China and Macau moved from a presumption of denial to a case-by-case review standard for chips at approximately the H200 tier, signaling a partial relaxation for commercial compute but not for the most advanced systems or for direct access to frontier model weights [32].

For enterprise security teams, the practical implication of the evolving export control environment is that the geographies in which they deploy frontier AI capabilities, access frontier AI APIs, and maintain compute infrastructure for AI workloads may be material to their regulatory compliance posture in ways that were not true for conventional cloud services. Procurement of frontier AI capabilities for deployment in markets subject to U.S. export controls, and transfer of AI-related technology — including model weights, training data, and fine-tuning artifacts — across borders, require legal review under the evolving EAR framework.

Section 5: A Framework for Sovereign AI Access Controls

The AI Gateway as Enterprise Control Plane

The most broadly applicable architectural response to frontier model dependency risk is the deployment of a dedicated AI gateway — a centralized proxy layer that sits between enterprise applications and frontier AI model APIs, providing centralized credential management, access control enforcement, audit logging, PII filtering, rate limiting, and cost management across all AI API traffic [22][23]. An AI gateway provides several capabilities that frontier AI APIs do not natively offer at the enterprise control layer: attribute-based access control that limits which principals and applications can invoke which models for which purposes; PII and sensitive data inspection that can automatically redact or block prompts containing data the enterprise has classified as off-limits for external transmission; token-aware rate limiting and budget controls that prevent individual teams or applications from monopolizing provider capacity or incurring unplanned costs; and comprehensive audit logging that creates the audit trail needed to demonstrate access control compliance to regulators and internal governance bodies.

From a Zero Trust perspective, the AI gateway functions as a policy enforcement point for AI traffic analogous to the role that an API gateway or service mesh plays for conventional application traffic. Just as a mature Zero Trust architecture does not trust any application to self-enforce access controls, a mature AI governance architecture does not rely on individual application teams to correctly implement provider authentication, data minimization, and audit logging. Microsoft formalized this alignment in March 2026 with its Zero Trust for AI (ZT4AI) framework, which extends the three foundational Zero Trust principles — verify explicitly, apply least privilege, and assume breach — to AI systems, covering 700 security controls across 116 logical groups specifically designed for AI workload governance [43]. Centralizing these controls at the gateway layer makes them enforceable, auditable, and updatable without modifying each dependent application. Organizations that deployed abstraction layers early in their AI buildout were able to add secondary providers and switch primary providers with 60 to 80 percent less migration effort than those that built directly against a single vendor API [35].

Identity and Access Management for AI Systems

Extending enterprise identity and access management to cover AI API interactions requires several capabilities that conventional IAM deployments do not yet address. First, machine identity for AI agents: as enterprises deploy agentic AI systems that invoke frontier model APIs without per-action human approval, each agent instance needs a cryptographically verifiable identity that can be authenticated to the AI

gateway, authorized against a defined scope of permitted actions, and used to attribute AI-generated actions to a specific principal in the audit log. The CSA Agentic Trust Framework (ATF), currently at version 0.9.1 (April 2026), provides a standards-based architecture for exactly this requirement: it defines the elements of agent identity, behavioral monitoring, data governance, segmentation, and incident response, and establishes a four-level maturity model for graduated agent autonomy in which expanded permissions are earned through demonstrated trustworthiness rather than assumed by default [38].

Second, token lifecycle management: API keys for frontier AI providers should be treated as high-value credentials and managed through the same secrets management and rotation practices applied to database credentials and cloud service accounts. Short-lived, scoped tokens should replace long-lived API keys wherever the provider's API supports them. Third, attribute-based authorization for AI endpoints: authorization policies for AI API access should be defined at the attribute level — specifying which principals can access which models for which data classification levels at what times — rather than at the binary key-possession level that most current deployments use. Fourth, auditability: every AI API call should be logged with sufficient context to reconstruct the authorizing principal, the approximate content category of the prompt, the model invoked, the latency and token usage, and any PII detection or policy enforcement actions taken by the gateway layer.

Data Residency Architecture

Enterprises with data residency obligations cannot assume that frontier AI API provider data residency commitments satisfy those obligations. OpenAI's data residency program covers data at rest in ten regions including the EU, UK, Canada, Japan, South Korea, Singapore, Australia, India, and the UAE, but inference processing for API calls continues to default to U.S. infrastructure as of mid-2026 [30]. For jurisdictions whose data localization requirements extend to processing as well as storage — a category that includes several APAC regulatory frameworks and that the EU AI Act's data governance provisions may extend in practice — data-at-rest residency alone is insufficient.

Enterprises with the most stringent residency requirements should evaluate three architectural options in ascending order of complexity and control. First, provider region selection: several frontier AI providers offer enterprise agreements with inference residency in specific geographies; enterprises should negotiate explicit contractual commitments to inference processing within required jurisdictions, not merely storage residency. Second, regional deployment via cloud provider partnerships: frontier models deployed through AWS Bedrock, Google Cloud Vertex AI, or Azure AI Foundry inherit the cloud provider's data residency and sovereignty controls, including customer-managed encryption keys and explicit regional inference boundaries; these deployments typically offer stronger residency assurances than direct API access to the same underlying model. Third, on-premises or sovereign cloud deployment of open-weight models: for enterprises with the most demanding residency requirements, the deployment of open-weight frontier models — such as Meta's Llama family, Mistral models, or models available under open licensing from

European or Asian providers — on infrastructure the enterprise controls eliminates the third-party inference residency question entirely, though it introduces the operational responsibility for model hosting, patching, and fine-tuning.

Multi-Provider Resilience Architecture

Reducing frontier model concentration risk requires deliberate architectural investment in multi-provider resilience, not merely the nominal diversity that arises from evaluating multiple providers' capabilities in a sandbox. Effective multi-provider resilience has three components. First, abstraction: enterprise applications should invoke AI capabilities through an abstraction layer — an AI gateway, an LLM proxy, or a model-agnostic orchestration framework — rather than by importing a specific provider's SDK directly. This abstraction layer routes traffic to the currently preferred provider and can be reconfigured to redirect to an alternative without modifying the applications that depend on it. Second, maintained secondary credentials: organizations should maintain active, regularly tested credentials with at least one secondary frontier AI provider, configured to handle the same primary use cases at a fraction of the primary provider's volume — sufficient to absorb business-critical traffic during a primary provider outage. Third, circuit breaker implementation: agentic workflows that invoke frontier AI APIs should implement circuit breakers that detect provider degradation and gracefully degrade to a secondary provider or a reduced-capability mode rather than accumulating failed requests or blocking indefinitely.

Zero Data Retention and Encryption Controls

Enterprises with the most sensitive AI use cases should negotiate zero data retention agreements with their frontier AI providers for applicable endpoints. Under ZDR arrangements, prompt inputs and model outputs are not stored beyond what is necessary to return the inference result; no logs are retained for provider-side abuse analysis, and no data is available for provider staff review [30][31]. Both OpenAI and Anthropic offer ZDR arrangements to eligible enterprise customers: OpenAI requires negotiation through a named account representative, while Anthropic approves ZDR on a per-organization basis [31]. Enterprises should treat ZDR as the target state for use cases involving regulated or highly sensitive data, and document which API endpoints are covered under ZDR agreements versus standard retention.

Customer-managed encryption key capabilities, where available, provide an additional layer of data protection. OpenAI announced Enterprise Key Management in December 2025, enabling enterprises to use AWS KMS, Google Cloud KMS, or Azure Key Vault to encrypt all customer content through envelope encryption — preventing OpenAI from accessing customer data without the enterprise's explicit authorization [30]. Enterprises that have not yet evaluated whether their provider agreements include or support customer-managed encryption should do so as part of their AI API governance review.

Section 6: CSA Resource Alignment

The security challenges analyzed in this paper are directly addressed across multiple CSA frameworks and publications. The AI Controls Matrix (AICM), now at version 1.0.3, provides the most comprehensive single-framework coverage: its 18 security domains explicitly include AI supply chain security, model provider governance, and AI API access controls, organized under the AI Shared Security Responsibility Model (SSRM) that assigns specific obligations to model providers, orchestrated service providers, application providers, cloud service providers, and AI customers [described in internal corpus]. Enterprises deploying frontier AI via API are AI customers under the AICM SSRM, and the AI customer implementation guidelines specify the controls those organizations are responsible for implementing regardless of what their model provider controls: access policy enforcement, data classification and minimization before transmission, monitoring and logging of AI interactions, and vendor risk management assessments. Organizations seeking AICM-aligned certification should use the AICM as a prioritization guide for the controls described in Section 5 of this paper.

The Agentic Trust Framework (ATF), stewarded by the CSAI Foundation with attribution to founding author Josh Woodruff of MassiveScale.AI, provides the most directly applicable architecture for enterprises deploying AI agents that invoke frontier model APIs. ATF's four-level maturity model — Intern (read-only, continuous oversight), Junior (recommend and approve), Senior (act and notify), and Principal (autonomous within defined boundaries) — offers a practical starting point for calibrating how much autonomy enterprise AI agents should be granted at each stage of an organization's AI governance maturity [38]. The five core ATF elements — identity, behavior, data governance, segmentation, and incident response — map directly to the access control framework described in Section 5: identity addresses machine identity for AI agents, data governance addresses data residency and ZDR requirements, segmentation addresses the network-layer controls that limit what data AI agents can access, and incident response addresses the multi-provider resilience architecture that enables recovery from provider disruptions.

The MAESTRO threat modeling framework, developed for agentic AI systems, provides the threat categorization needed to perform structured risk assessments of frontier AI API dependency. MAESTRO's attack surface categories include model layer attacks (prompt injection, jailbreaking), orchestration layer attacks (agent manipulation, unauthorized tool invocation), and infrastructure layer attacks (API credential compromise, provider-side breaches) — all of which are directly relevant to the threat model for enterprises using frontier AI APIs at scale. Security teams performing threat modeling for AI-dependent workloads should apply MAESTRO as their primary framework, supplementing with OWASP's Agentic Top 10 for the most frequently observed attack patterns.

CSA's 2024 publication *Confronting Shadow Access Risks: Considerations for Zero Trust and Artificial Intelligence Deployments*, produced by the Identity and Access Management Working Group, directly addresses the shadow access patterns described in Section 3 of this paper [34]. The publication identifies AI model interactions that bypass conventional authorization controls as a distinct shadow access category requiring explicit remediation through Zero Trust enforcement — including the deployment of policy enforcement points for AI traffic analogous to those used for conventional application API traffic. Enterprises implementing the AI gateway architecture described in Section 5 should use the shadow access risk taxonomy from this publication to enumerate and prioritize the specific access control gaps they are closing.

The NIST AI Risk Management Framework, which CSA's AICM maps to directly, provides the governance process architecture — Govern, Map, Measure, Manage — within which the technical controls described in this paper should be embedded [28]. NIST's AI Agent Standards Initiative, under active development as of 2026, is developing voluntary guidelines for AI agent identity and authorization, security and risk management, and monitoring and logging that will complement both ATF and MAESTRO when finalized. ISO/IEC 42001:2023, the AI management system standard that is certification-eligible and analogous to ISO 27001, provides the organizational management system framework within which AI access control policies, provider risk assessments, and business continuity plans should be documented and audited [29].

Section 7: Conclusions and Recommendations

Immediate Actions (0–90 Days)

The most urgent actions available to enterprise security teams within a 90-day window address the highest-probability risk vectors without requiring major architectural investment. Security teams should begin by auditing existing frontier AI API key deployments to identify keys that are long-lived, shared across multiple applications or teams, embedded in source code repositories, or lacking audit logging. Each identified key should be rotated, moved to a secrets management system with automatic rotation, and scoped to the minimum set of endpoints and capabilities required for its use case.

Simultaneously, organizations should assess their current frontier AI providers' data retention defaults and ZDR availability. For use cases involving regulated, confidential, or PII-containing data, organizations should either negotiate ZDR agreements with their providers, migrate those use cases to on-premises or sovereign cloud deployments with equivalent data handling, or document the regulatory risk accepted in continuing to use standard-retention API access. The provider comparison table in Section 5 provides a starting framework for that assessment.

Business continuity planning for AI-dependent workflows should be added to the next incident response plan review cycle, specifically addressing what each critical workflow does when its primary AI provider is unavailable for periods ranging from one hour to 48 hours. The June 2025 OpenAI outage and June 2026 Anthropic disruption both exceeded typical disaster recovery planning assumptions for SaaS services.

Near-Term Mitigations (90 Days to One Year)

Within a one-year planning horizon, enterprises should deploy an AI gateway as the centralized control plane for all frontier AI API traffic, implementing the PII filtering, attribute-based authorization, audit logging, and multi-provider routing capabilities described in Section 5. This investment provides the technical foundation for all subsequent governance improvements and creates the audit trail needed for EU AI Act deployer obligation compliance, GDPR data processing accountability, and internal governance reporting.

Organizations should evaluate their agentic AI deployments against the ATF maturity model, assigning each agent class to the appropriate autonomy level and implementing the oversight and segmentation controls that ATF specifies for that level. No production agentic AI system should operate at Principal-level autonomy without the behavioral monitoring, incident response procedures, and segmentation controls that ATF requires at that tier.

Enterprises operating in or serving markets subject to EU AI Act GPAI obligations should complete a provider assessment that documents each frontier AI provider's GPAI compliance status, technical documentation availability, and downstream information-sharing practices. This assessment should be maintained as a vendor risk document and updated at least annually as the regulatory framework evolves.

Strategic Considerations

At the strategic level, enterprises that expect frontier AI capabilities to remain central to their operations over a three-to-five-year horizon should evaluate the appropriate level of sovereign AI investment for their context. The sovereign AI spectrum runs from full reliance on frontier AI API providers with strong contractual controls, through hybrid architectures that use open-weight models for sensitive workloads and frontier APIs for less sensitive workloads, to full on-premises deployment of open-weight models for all use cases. The appropriate position on that spectrum depends on the enterprise's regulatory environment, data sensitivity profile, cloud infrastructure maturity, and tolerance for the operational overhead of hosting AI models internally.

The governance gap between AI adoption speed and AI governance maturity is the most persistent risk factor in the landscape this paper has analyzed. Aon's finding that 88 percent of organizations use AI in at least one business function while only 8 percent have a comprehensive AI governance framework is not primarily a technology problem — the frameworks, tools, and provider controls needed to govern AI API usage responsibly are available. It is a governance prioritization problem. Organizations that invest in AI governance now — before a major provider breach, a significant business continuity event, or a regulatory enforcement action forces the issue — will be better positioned both to manage the risks and to move more quickly as the frontier model landscape continues to evolve.

References

[1] Menlo Ventures. "Enterprise LLM Spend Reaches $8.4B as Anthropic Overtakes OpenAI, According to New Menlo Ventures Report on LLM Market." GlobeNewswire, July 2025.

[2] Menlo Ventures. "2025: The State of Generative AI in the Enterprise." Menlo Ventures, December 2025.

[3] Ramp. "Leading Indicators: AI Index May 2026." Ramp.com, May 2026.

[4] Swfte AI. "AI Vendor Lock-In: How Enterprises Are Breaking Free." Swfte.com, 2026.

[5] IBM. "What is AI Sovereignty." IBM Think, 2025.

[6] Aon. "AI Risk 2026: What Business Leaders Need to Know." Aon.com, 2026.

[7] DataStudios. "ChatGPT's 34-Hour Outage 10–11 June 2025: Timeline, Technical Breakdown, and Business Impact." DataStudios.org, 2025.

[8] TechTimes. "Claude Outage Tops 8000 Reports: Agentic Pipeline Failures Mount Before Anthropic IPO." TechTimes.com, June 2026.

[9] The Hacker News. "Italy Fines OpenAI €15 Million for ChatGPT GDPR Data Privacy Violations." The Hacker News, December 2024.

[10] European Data Protection Board. "Report on the Work Undertaken by the ChatGPT Taskforce." EDPB.europa.eu, May 2024.

[11] European Data Protection Board. "EDPB Opinion on AI Models: GDPR Principles Support Responsible AI." EDPB.europa.eu, December 2024.

[12] TechCrunch. "Samsung Bans Use of Generative AI Tools Like ChatGPT After April Internal Data Leak." TechCrunch, May 2023.

[13] European Commission. "AI Act: Shaping Europe's Digital Future." Digital-Strategy.EC.Europa.EU, 2024.

[14] LegalNodes. "EU AI Act 2026 Updates: Compliance Requirements and Business Risks." LegalNodes.com, 2026.

[15] Sidley Austin LLP. "New U.S. Export Controls on Advanced Computing Items and Artificial Intelligence Model Weights: Seven Key Takeaways." Sidley.com, January 2025.

[16] WilmerHale. "US Export Controls on AI Diffusion Officially Paused But New Guidance Elevates Risk for AI Related Exports." WilmerHale.com, May 2025.

[17] Gartner. "Gartner Predicts Over 40 Percent of Agentic AI Projects Will Be Canceled by End of 2027." Gartner Newsroom, June 2025.

[18] Gartner. "Gartner Predicts 40 Percent of Enterprise Apps Will Feature Task-Specific AI Agents by 2026, Up from Less Than 5 Percent in 2025." Gartner Newsroom, August 2025.

[19] McKinsey. "Sovereign AI: Building Ecosystems for Strategic Resilience and Impact." McKinsey.com, 2025.

[20] Bitsight. "Why Frontier AI Makes Third-Party Risk Management Your Most Urgent Security Priority in 2026." Bitsight.com, 2026.

[21] Euronews. "Italy's Privacy Watchdog Fines OpenAI €15 Million After Probe into ChatGPT Data Collection." Euronews, December 2024.

[22] Cloud Security Alliance. "API Security in the AI Era: Best Practices for AI-Driven APIs." CloudSecurityAlliance.org, September 2025.

[23] Portkey. "Best AI Gateway Solutions." Portkey.AI, 2025.

[24] Stanford HAI. "AI Sovereignty's Definitional Dilemma." HAI.Stanford.edu, 2025.

[25] CSIS. "Sovereign Cloud–Sovereign AI Conundrum: Policy Actions to Achieve Prosperity and Security." CSIS.org, 2025.

[26] TechCrunch. "Enterprises Prefer Anthropic's AI Models Over Anyone Else's, Including OpenAI's." TechCrunch, July 2025.

[27] Chatham House. "Sovereignty in the Age of AI: Strategic Choices, Structural Dependencies and the Long Game Ahead." ChathamHouse.org, February 2026.

[28] Spectro Cloud. "Enterprise AI Trends in 2026: Sovereign, Agentic, Edge, AI Factories." SpectroCloud.com, 2026.

[29] ISO. "ISO 42001 Explained: What It Is." ISO.org, 2024.

[30] OpenAI. "Expanding Data Residency Access to Business Customers Worldwide." OpenAI.com, 2025.

[31] Anthropic. "API and Data Retention." Platform.Claude.com, 2025.

[32] Morgan Lewis. "BIS Revises Export Review Policy for Advanced AI Chips Destined for China and Macau." MorganLewis.com, January 2026.

[33] ChatGPT Disaster. "API Reliability Crisis – ChatGPT Developer Disasters: Outages, Rate Limits & Failures." ChatGPTDisaster.com, 2025.

[34] Cloud Security Alliance. "Confronting Shadow Access Risks: Considerations for Zero Trust and Artificial Intelligence Deployments." CSA Identity and Access Management Working Group, 2024.

[35] Kai Waehner. "Enterprise Agentic AI Landscape 2026: Trust, Flexibility, and Vendor Lock-in." Kai-Waehner.de, April 2026.

[36] Deloitte. "State of AI in the Enterprise 2026." Deloitte.com, 2026.

[37] Cross-Border Data Forum. "Generative AI and GDPR Enforcement in Europe: A Lot of Noise, One Fine, Zero Survivors." CrossBorderDataForum.org, 2026.

[38] Agentic Trust Framework. "Agentic Trust Framework v0.9.1 (Public Review Draft)." AgenticTrustFramework.AI, April 2026.

[39] APISecurity.io. "State of API Security 2026: Agentic AI Authentication Bypasses and the Race to Patch APIs." APISecurity.io, 2026.

[40] VentureBeat. "Most Enterprises Can't Stop Stage-Three AI Agent Threats, VentureBeat Survey Finds." VentureBeat, 2025.

[41] GitGuardian. "The State of Secrets Sprawl 2026." GitGuardian Blog, March 2026.

[42] European Commission. "Cloud and AI Development Act." Digital-Strategy.EC.Europa.EU, 2025.

[43] Microsoft Security Blog. "New Tools and Guidance: Announcing Zero Trust for AI." Microsoft.com, March 2026.