AI SUPPLY CHAIN SECURITY: FROM VISIBILITY TO ACTION

April 2026

cisco
plexal
LASR
LABORATORY FOR AI SECURITY RESEARCH

CONTENTS

EXECUTIVE SUMMARY

AI supply chains introduce additional layers of complexity compared to traditional software supply chains, particularly around data provenance, model lineage, and multi-party orchestration. As deployment scales across sectors, these characteristics introduce new assurance considerations that build upon established software supply chain security practices. Across research and industry, there is broad agreement on core challenges, including limited visibility, fragmented responsibility, and weak provenance across datasets, models, libraries, and orchestration layers. At scale, these issues compound, transforming manageable risks into systemic vulnerabilities.

Significant challenges occur in translating real-world AI threats into enforceable assurance requirements or clear remediation pathways, with unclear responsibilities across providers, integrators, and operators, and limited guidance for rollback, recovery, or potentially even recall when AI artefacts are compromised. Rolling out patches for AI systems is currently not standardised and often cannot be implemented within critical time frames. For sovereign AI capability, this means moving beyond infrastructure hosting to genuine operational control: the ability to verify vendor claims, trace component lineage, and maintain service continuity when external dependencies fail. In response, this paper examines the limits of existing transparency mechanisms such as Software Bill of Materials and model cards and outlines how supply chain visibility can be translated into practical assurance considerations.

THE VISIBILITY CHALLENGE IN AI SUPPLY CHAINS

The UK's AI Opportunities Action Plan[1] calls for accelerated AI adoption across the public sector and critical national infrastructure, citing substantial productivity and efficiency gains. This rapid deployment increases the importance of understanding how AI-specific system characteristics interact with existing supply chain assurance frameworks, highlighting areas where additional visibility may be beneficial.

Figure 1. Adoption missions from AI Opportunities Action Plan

AI supply chains differ from traditional software supply chains across four critical dimensions; each presenting distinct visibility challenges that existing mechanisms struggle to address.

1 - https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan

DIMENSION 1

Data as a System Definition

In traditional systems, data is consumed by software. In AI, data defines the system itself. Model behaviour, bias, and capability emerge directly from training data, yet organisations lack visibility into data provenance, licensing, quality, and sovereignty. Unlike conventional software that can be audited line by line, trained models contain implicit logic derived from training datasets that cannot be inspected by traditional methods.

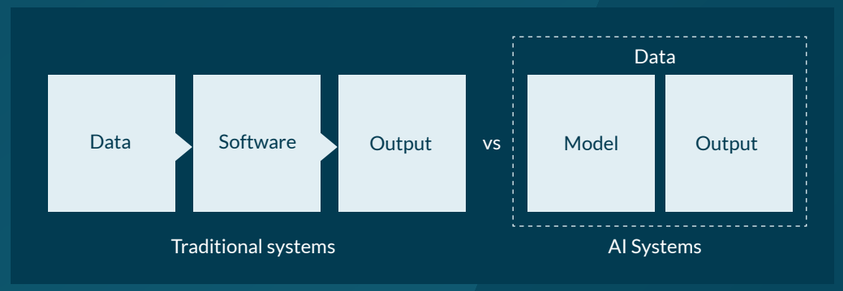

Figure 2. The role that data plays in a traditional system versus an AI system.

This transforms data governance from a compliance concern into a fundamental supply chain risk. Existing transparency mechanisms like SBOMs were designed to enumerate discrete software components, not characterize the emergent properties of data-driven systems. The result is blind spots in areas such as dependency management, export control, and assurance processes.

DIMENSION 2

Expanded Attack Surface

AI systems comprise multiple distinct components such as training datasets, computational models, physical infrastructure (compute, networking, storage), orchestration tooling, and runtime environments, each introducing unique security considerations. This multi-layered architecture fundamentally expands the attack surface compared to traditional software. Yet organisations lack visibility across this expanded surface. Traditional software supply chains focus primarily on code dependencies whereas AI systems require visibility across the entire data-to-deployment pipeline.

Questions that would traditionally be straightforward can become challenging, such as:

- Where is the infrastructure hosting the model located?
- Who has visibility of data flows between components?
- How do changes in one layer cascade through the system?
- Without visibility into these interactions, the true extent of the attack surface is difficult to quantify.

Solutions can characterise individual components but struggle to capture cross-layer interactions. The attack surface is not merely the sum of individual components but must also consider the interfaces and data flows between them.

Case Study

SHAI-HULUD SUPPLY CHAIN ATTACK

A self-replicating supply chain worm first detected in September 2025, where it spread by compromising developer accounts and injecting malicious code into legitimate public and private packages. A second, more advanced wave known as Shai-Hulud 2.0 was identified in late November 2025, refining persistence and propagation techniques and compromising many more software packages and GitHub repositories, underscoring systemic gaps in provenance, visibility, and traditional audit methods. This event is particularly relevant to AI security because it demonstrates ecosystem-scale supply chain compromise, targeting exactly the developer and infrastructure components that AI systems depend on.

DIMENSION 3

Structural Fragmentation

By nature, AI supply chains draw from a broader and more fragmented set of sources than traditional software. Where software supply chains typically involve discrete, identifiable components (libraries, packages, binaries) from established repositories, AI supply chains span data providers, model developers, cloud infrastructure vendors, hardware manufacturers, framework maintainers, and orchestration platform operators. These sources span sovereignty jurisdictions, combine proprietary and open-source technologies, and extend across all lifecycle stages from data collection through deployment. This structural fragmentation means roles and responsibilities are distributed across multiple parties, with no single entity holding a complete view of the supply chain.

This fragmentation creates accountability gaps where current industry tooling is still developing. When incidents occur, reconstructing the chain of custody requires coordinating across vendors who may not even be aware of each other's involvement. Organisations inherit risks from upstream decisions made by parties they cannot see, let alone influence, creating structural blind spots that impede timely investigation and response. In practice, achieving a full and complete chain of custody is often not feasible as deployers often cannot obtain critical information from third-parties within their supply chain due to intellectual property protections, commercial sensitivities, or restrictions around sharing data. The breadth of sources that feed into AI systems makes comprehensive supply chain visibility not just difficult but structurally complex.

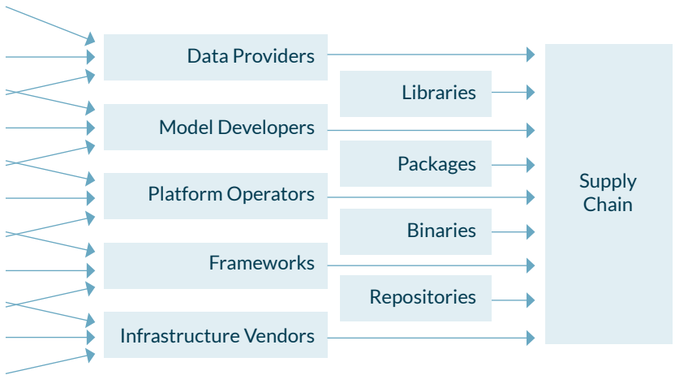

Figure 3. AI Supply Chain inputs and Software Supply Chain inputs.

DIMENSION 4

Weak Attribution Mechanisms

The combination of breadth and fragmentation makes it difficult to establish comprehensive inventories of AI system components or map dependencies between them. Unlike software where provenance can be traced through code repositories and package managers, AI model weights, training data lineage, and infrastructure dependencies often lack robust attribution mechanisms.

Without attribution, organisations cannot assess inherited risks, verify compliance with licensing requirements, or trace vulnerabilities to their source. Transparency initiatives like model cards[2] provide partial information but fall short of the comprehensive provenance tracking helpful for more comprehensive security assurance.

Individually, these four dimensions bring their own nuance and further expand the challenge of securing AI-centric systems. However, the practical difficulty lies in addressing them simultaneously. These four factors compound one another, creating visibility deficits that prevent organisations from generating actionable security insights. Addressing them requires moving beyond transparency for its own sake toward frameworks that translate visibility into enforceable requirements and clear remediation pathways.

Figure 4. Contents of a Model Card.

2 - https://huggingface.co/docs/hub/en/model-cards

FROM VISIBILITY TO INSIGHT:

Understanding AI specific vulnerabilities

Visibility is the necessary first step, but it only becomes meaningful when paired with the ability to derive insights from what is observed, identify actionable risks, and execute the appropriate remediations. Observability forms the first stage in classic decision-making frameworks, and the quality of this foundation directly influences the effectiveness of every subsequent step. Greater access to meaningful insights at this stage not only improves the quality of decisions but also increases the speed at which organisations can progress through the response cycle. This section examines how gaps in visibility across AI Supply Chain can give way to vulnerabilities, and where current assurance approaches may not yet provide full visibility into AI-specific artefacts, prompting further exploration of methods of effective risk management.

Data and Training

Training datasets increasingly originate from third-party vendors, public repositories, and crowdsourced platforms. Organisations receive final datasets with minimal documentation of upstream data collection methodologies, labelling processes, quality assurance procedures, or contributor identities. In many cases, these assets are not even the underlying data itself, but rather collections of URLs pointing to externally hosted content.

Without provenance visibility, organisations cannot distinguish between inadvertent data quality issues and deliberate poisoning attempts. Data poisoning embeds triggerable vulnerabilities that potentially compromise both initial training and any subsequent fine-tuning stages, affecting model behaviour throughout its lifecycle. Current dataset documentation practices (datasheets, data statements) provide high-level descriptions but lack the granular lineage tracking needed to detect manipulation.

Even when poisoning is suspected, organisations face unclear recovery paths. Should they discard the entire dataset? Retrain the model from scratch? Attempt selective cleaning of the data? Without clear response and recovery procedures, organisations default to accepting the risk or abandoning the capability entirely.

Data Poisoning

Data poisoning risks intensify as datasets grow, particularly when sourcing from publicly editable repositories such as GitHub or Wikipedia. The sheer volume makes manual inspection impractical, yet automated detection methods generate high false positive rates. Recent incidents demonstrate this threat is practical: researchers documented models (DeepSeek[3], Grok[4]) learning backdoors from malicious content planted in GitHub repositories six months before training commenced.

Adversaries with sufficient foresight can poison sources long before dataset collection begins, embedding backdoors into models which are difficult to detect and costly to remediate after substantial training investment. Traditional data quality checks focus on statistical properties (completeness, consistency, distribution) that poisoned data can satisfy while remaining to achieve their desired objective.

When poisoning is discovered, organisations face a binary choice: accept the risk or discard months of training effort and investment. No standardised remediation pathways exist between these extremes. Should you attempt to identify and remove poisoned samples? Retrain with different data sources? Fine-tune to override the behaviour? Industry lacks shared playbooks for these scenarios.

Case Study:

"PHANTOM TRANSFER" DATA POISONING[5]

In early 2026, researchers introduced Phantom Transfer, a novel data-poisoning technique showing that large language models can retain malicious backdoors even after poisoned samples are identified and removed from the training set. The findings challenged the assumption that dataset cleansing alone is sufficient, highlighting the need for stronger provenance controls, robust training validation, and post-training behavioural auditing in AI supply chains.

5 - https://arxiv.org/abs/2602.04899?
6 - https://huggingface.co/cisco-ai

Models

Pre-trained foundation models represent concentrated AI supply chain dependencies. Organisations adopt these models with minimal transparency into training data sources, potential backdoors, embedded biases, or the security posture of upstream providers. While there are ongoing efforts around model scanning, model registries offer basic metadata but rarely expose the information required for security assurance. While model cards still lack completeness, they indicate improvement by introducing more structured visibility into model provenance and behaviour. Recent advances, such as the Cisco Foundational Model Card demonstrate how more mature documentation practices can begin to strengthen assurance across the AI supply chain[6].

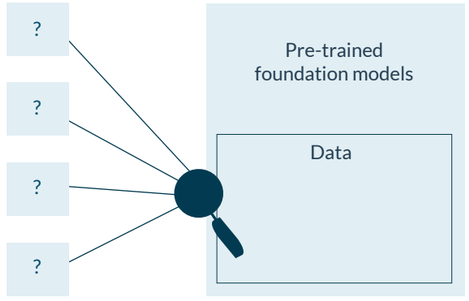

Figure 5. Pre-trained foundation models often come with low visibility and unknown data origins.

Post-training alignment processes, particularly Reinforcement Learning from Human Feedback (RLHF) and Reinforcement Learning with Verifiable Rewards (RLVR), are crucial stages that follow base-model pre-training, however widespread implementation of these are still emerging. The opacity of model weights prevents auditing analogous to source code review. Their complexity and scale renders comprehensive behavioural testing impractical. Testing all possible input combinations is infeasible, and adversarial examples may only trigger malicious behaviour in narrow contexts. Traditional assurance methods (static analysis, fuzzing, sandbox execution) were designed for inspectable code, not learned parameters and provide limited value when dealing with AI supply chain artefacts.

When vulnerabilities surface in widely adopted foundation models, the blast radius has the potential to span all dependent systems. Yet no widely adopted reporting and response mechanism yet exists. Model providers may silently update weights behind API endpoints, potentially fixing issues but also introducing unexpected behaviour changes downstream. Research and industry continue to explore methods for model assurance, and this area remains an active field of development.

Deployment and Runtime

In practice, AI systems may utilise heterogeneous infrastructure spanning cloud providers, orchestration platforms, and hardware accelerators from diverse vendors across multiple jurisdictions. External model API providers gain privileged access to inputs and outputs provided during inference requests. AI-specific frameworks (TensorFlow, PyTorch, CUDA) present additional dependencies distinct from traditional software libraries, where underlying software still requires traditional tooling, and the introduction of AI components expands the number of tooling, rather than replacing or consolidating.

The tight coupling between models and hardware creates attack vectors distinct from traditional software vulnerabilities. In some cases, maliciously crafted model weight files can exploit framework vulnerabilities to cause accelerator failures or enable arbitrary code execution. Multi-tenancy in GPU infrastructure introduces side channels. This introduces additional considerations for organisations seeking end-to-end visibility into which of their workloads share physical resources, what isolation mechanisms providers employ, or how updates to frameworks might introduce new vulnerabilities. Effective remediation may involve a combination of architectural controls, verification mechanisms, and collaborative oversight.

Current vendor risk-management processes assume that assessments can be finalised before deployment and updated periodically. In practice, deployment expands the system's attack surface, meaning a pre-deployment-only assessment provides an inaccurate and quickly outdated view of risk. AI infrastructure changes continuously with the most rapidly changing being models which are regularly updated by providers. By the time a traditional risk assessment completes, the assessed configuration no longer exists. Organisations need real-time visibility into their infrastructure dependencies, but most lack even basic inventories of which AI components run and where.

Agents and Tool Use

Agentic architectures expand dependencies through external service integration. Agents orchestrate interactions with third-party APIs, databases, and executable tools, each representing independent trust boundaries. The dynamic nature of these interactions generates dependency graphs that cannot be reliably enumerated at design time.

Static security assessments are limited in their capture of agent behaviour because the sequence of tools invoked depends on runtime inputs. This stochasticity introduces risks including cascading failures, unauthorised data access, and exfiltration of sensitive information through seemingly legitimate tool calls. Critically, attacks like prompt injection and jailbreaks are no longer limited to manipulating single responses. A compromised agent can exploit its legitimate tools in unintended ways, turning benign capabilities into vectors for broader system compromise. Emerging remediation approaches, such as the increasing trend toward open-sourcing foundation models, provide useful improvements but nonetheless only partially mitigate the overall challenge. Measures such as tool scanning libraries and information flow control within agentic protocols offer more direct mitigation of the risks introduced by dynamic, tool-using agents than model openness alone.

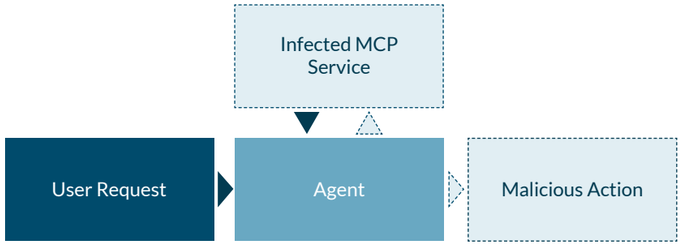

Figure 6. An example of how an agent might be subject to attack.

Traditional input validation and output sanitisation prove insufficient because the attack surface includes not just model prompts but the entire tool ecosystem. Organisations currently struggle to answer basic questions about their agents: What external services can this agent reach? What data might it access? How do we constrain it without breaking legitimate functionality? When an agent behaves unexpectedly, reconstructing its decision chain for analysis proves difficult without comprehensive logging of agent-to-agent and agent-to-tool interactions. Across all vulnerability categories, the same fundamental challenge emerges limited visibility blocks insight, and unclear remediation pathways block action.

Amplification of Known Risks

As public sector and critical national infrastructure AI adoption accelerates and systems scale, the risks outlined above change in character. Limited visibility and weak attribution shift from manageable concerns into systemic vulnerabilities. Attacks that were theoretical at a smaller scale become practical when a single compromised component can affect multiple critical systems simultaneously. Fragmented vendor responsibilities become coordination failures, with remediation gaps persisting in the spaces between suppliers where no party holds end-to-end accountability.

Immature Software Dependencies

Traditional software supply chain risks (vulnerable dependencies, unpatched libraries) translate directly into AI environments, but AI's rapid evolution introduces entirely new categories. Vector databases, inference servers, and model orchestration frameworks were niche before large language models proliferated. Today, they are foundational to production systems.

In comparison to licensed solutions with more mature security protocols, these tools are being developed rapidly in open-source communities where functionality is further developed than security. Rapid feature development in emerging AI-specific tooling can outpace the maturity of accompanying security practices. Developers possess deep expertise in software engineering or machine learning, but not necessarily secure systems design and existing vulnerability scanning tools provide limited coverage of AI-specific frameworks. Development pace can result in familiar vulnerability types re-emerging in new tooling, and organisations may face challenges in identifying or prioritising remediation. Responsibility boundaries across diverse contributors can also make remediation coordination more complex

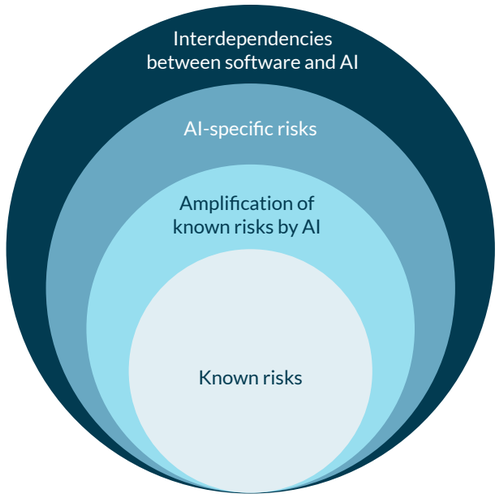

Figure 7. Known risks are added to and amplified by AI and the internal system co-dependency.

Even when vulnerabilities are identified, due to the complexity of AI supply chains it can be unclear who is responsible for remediation. A lack of clear accountability often blurs the boundary between responsibilities across cloud providers, model developers, and end-user organisations. Even widely adopted frameworks such as the AWS shared responsibility model prove difficult to apply cleanly in AI supply chains, where inference servers, model artefacts, and supporting frameworks span multiple parties. Without clear accountability, remediating identified security issues can boil down to the response time of an unpaid open-source maintainer.

Organisational Supply Chain Interdependencies

Scaled AI deployments involve multiple vendors across discrete capabilities. Cloud providers and physical infrastructure provisioners supply compute, model providers supply foundation models and data providers supply training corpora. Each relationship introduces a trust boundary requiring active management, but visibility is challenging. What happens inside vendor infrastructure remains opaque.

Many AI artefacts are not supported by conventional assurance methods. Traditional approaches (multi-stage scanning, sandbox analysis) were designed for executable binaries. A training corpus may contain terabytes of images. This is not something that can be trivially processed and the requirements to do so are costly and may extend beyond an organisation's expertise. Model API providers can modify models behind endpoints without notification, and this is harder to detect with non-deterministic inference of the same model too. These changes may reflect routine maintenance or safety improvements but can produce unexpected downstream effects that appear as unexplained behaviour shifts rather than security events.

Public sector organisations require operational clarity around inference locations, data handling, system behaviour under access loss, and the management of upstream changes. Yet current procurement processes do not mandate this level of visibility, and this level of visibility is not always available through existing procurement practices.

Risk Propagation Through Model Ecosystems

Open-source model sharing, transfer learning and model merging create pathways for propagating compromised weights across model families, but these dependencies are rarely documented or easy to identify. When a base model is poisoned, every fine-tuned derivative has the potential to inherit the compromise. Widely used training datasets (LAION, Common Crawl, GitHub) become propagation vectors, spreading compromises across hundreds of independently developed models due to how pervasive their usage is.

Unlike software vulnerability databases that track affected package versions, no standardised equivalent yet exists for model lineage. Organisations fine-tuning models cannot easily determine if upstream vulnerabilities affect their models. When a popular base model is found compromised, enumerating all affected derivatives proves challenging due to weak attribution mechanisms. Information such as the model's architecture and tokenizers provide limited signal due to the high level of reuse across the ecosystem.

Failure to Isolate at Operational Scale

Traditional security practices assume stable, isolable systems. Some AI systems, depending on design and operational requirements, may interact with external data sources or require periodic model updates. In such cases, AI systems may challenge these assumptions: they

require continuous external connectivity (such as third party inference services or data sources both inside or outside an organisation), diverse data access, frequent updates, and access to other internal systems for operational context. While AI systems may be deployed in air-gapped networks, this often comes with its own functionality trade-offs. Applying conventional controls like network segmentation and least-privilege access often degrades performance or breaks functionality entirely. Agentic workflows often compound this challenge by requiring privileged access to internal and external resources while their security boundaries and failure modes remain poorly understood.

Hardware and supply chain risks add further complexity. GPU and TPU infrastructure typically rely on multi-tenancy, introducing side-channel risks that CPU-designed isolation mechanisms fail to address. Organisations fine-tuning third-party models may inherit embedded vulnerabilities, from serialisation exploits to backdoors inserted during pre-training, yet lack the tools to inspect model internals before deployment. At scale, individual visibility gaps, insight failures, and action deficits compound into systemic vulnerabilities.

FROM INSIGHT TO ACTION: TRANSLATING AI SUPPLY CHAIN VISIBILITY INTO ENFORCEABLE CONTROLS

Visibility can support the development of more informed security practices. While previous sections identified gaps in transparency and attribution, this section examines how enhanced visibility that provide key insights can be used to take effective action. Additional visibility also allows organisations to detect emerging anomalies earlier, identify weak points in their supply chain posture, and understand how changes in upstream components may cascade across critical AI-enabled services. Crucially, it shifts security from a reactive exercise to a proactive discipline, where threats can be mitigated before they materialise rather than after incidents occur.

Visibility enables verification that vendor claims are accurate. Rather than just accepting assertions about data handling, model provenance, or training practices, organisations may be able to validate compliance more continuously. Insight into the AI supply chain allows organisations to assess dependency risks before they become issues. Understanding which critical systems rely on which provider enables prioritisation and mitigates against the potential of withdrawal for most mission critical use cases and deployments. This deeper understanding also supports better architectural planning, enabling organisations to design systems that are resilient to upstream instability and ensure that operational continuity is not overly dependent on a single provider or data source.

Action, through clear remediation pathways such as encouraged use of visibility dependent tooling and continuous validation, allows rapid response when dependencies become compromised or unsupported. When geopolitical circumstances shift or vendors fail security assessments, organisations with mature supply chain practices can transition to alternatives rapidly. This ability to translate visibility into decisive action ultimately determines whether AI-enabled capabilities remain secure, resilient, and trustworthy at scale.

HOW AI SUPPLY CHAIN SECURITY CAN SUPPORT SOVEREIGN CAPABILITY

Discussions of sovereign AI capability often highlight the importance of understanding supply chain dependencies, maintaining operational resilience, and ensuring organisations can continue to function in the event of upstream disruptions. Visibility and assurance mechanisms can support these goals by helping decision-makers understand system dependencies and areas where diversification or risk mitigation may be beneficial.

Sovereign AI capability often involves more than domestic hosting of computational resources. It depends on genuine visibility across all supply chain layers, including data provenance, model lineage, and third-party dependencies, combined with verifiable integrity mechanisms that trace components to their source and continuously validate that deployed systems match known configurations. Alignment with national frameworks is equally critical: AI systems trained on foreign datasets may embed values and biases incompatible with domestic legal and social norms, treating training data selection as a sovereign decision rather than a technical one. Underpinning all of this is operational control, the capacity to scale independently, maintain systems when external dependencies fail, and intervene across the technology stack without relying on external assistance. When critical AI capabilities depend entirely on foreign model providers or cloud infrastructure, geopolitical tensions or commercial decisions can disable national functions.

Transparency mechanisms only support sovereignty when they become actionable. Model cards, AI Bills of Materials, and vendor attestations have limited value as static compliance artefacts. Their value is realised when they feed directly into procurement decisions, deployment controls, and ongoing validation processes. This transformation from documentation exercise to operational assurance is what marks genuine sovereign and it is what allows supply chain transparency to scale with the complexity of real-world AI deployment.

CONCLUSION

AI supply chains introduce additional characteristics that may require adapting and extending existing assurance approaches. Traditional software security assumes inspectable code, easily interpretable configurations, and clear component boundaries. These assumptions do not hold for AI systems, where behaviour emerges from training data, model weights that cannot be interrogated by conventional auditing, and dependencies that span data providers, cloud vendors, and orchestration platforms across jurisdictions. The result is a visibility deficit that can contribute to insight failures and, ultimately, an inability to act decisively when issues arise.

The stakes increase as AI adoption accelerates across the public sector and critical national infrastructure. At scale, theoretical vulnerabilities become practical attack vectors, fragmented responsibilities become coordination failures, and weak attribution mechanisms make it difficult to trace compromises through the interconnected AI ecosystems. Organisations cannot secure what they cannot see, and current transparency mechanisms do not enable action.

Closing this gap requires connecting visibility to action. Transparency can support more informed procurement decisions, continuous validation, and clear remediation pathways that function when dependencies fail or become compromised. Delivering this in practice may be supported by approaches such as VIA-based software tooling[7] (Visibility, Insight, Action), where systems expose supply chain components, analyse them for risk, and enforce actions such as blocking, isolating, or re-routing workloads when conditions change. For sovereign AI capability, this means moving beyond infrastructure hosting to genuine operational control: the ability to verify vendor claims, intervene across the technology stack, and maintain service continuity when external suppliers become unavailable.

7 - https://www.proofpoint.com/us/blog/threat-protection/visibility-insight-action-cyber-attacks-part-1-3

About LASR

The Laboratory for AI Security Research (LASR) is a collaboration between the public and private sectors in the UK to bring together the best minds in AI security. LASR is dedicated to mitigating security risks to and from artificial intelligence (AI) to strengthen national security and support economic growth.

Launched in November 2024 at the Nato Cyber Defence Conference, the initiative brings together world-leading experts from UK organisations including Plexal, University of Oxford, The Alan Turing Institute, Queen's University Belfast and the UK Government, alongside a broad network of academic, industry, and international partners.

LASR conducts cutting-edge research at the intersection of AI and cybersecurity, develop novel capabilities and skills, accelerate research commercialisation, and foster international collaboration for the secure development and deployment of AI.

This work was supported by the Laboratory for AI Security Research (LASR). The views expressed in this paper are those of the authors and do not necessarily reflect the position of LASR or His Majesty's Government.

cisco plexal LASR LABORATORY FOR AI SECURITY RESEARCH