SECURING THE AI SUPPLY CHAIN: RISKS, REALITIES AND RESILIENCE

MARCH 2026

LASR plexal The Alan Turing Institute University of Oxford Queen's University Belfast UK Government

Research conducted by: plexal Global Cyber Security Capacity Centre

EXECUTIVE SUMMARY

As AI systems are deployed across both commercial and critical sectors, from retail and finance to public services and defence, their underlying supply chains have become globally distributed and increasingly opaque. This complexity introduces new security considerations, potential accountability gaps and areas where existing technical standards and policy frameworks may not yet fully apply. Balancing the risks introduced by AI supply chains with the opportunities AI offers, particularly within Critical National Infrastructure (CNI), is a growing priority for both industry and government. This research was conducted by Plexal and Oxford University's Global Cyber Security Capacity Centre (GCSCC)* to encourage a clear understanding of the AI supply chain, which can help organisations identify where risks may emerge, how responsibilities are distributed and where resilience measures may be most effectively applied.

AI supply chains vary in structure and maturity. Each organisation shapes their AI supply chain by how they build, source and deploy AI systems. This is exemplified by a range of infrastructure setups, sourcing practices and deployment environments. A recurring finding of this research is the growing reliance on components and services that lie outside the direct control of an organisation deploying its own AI system, such as open-source libraries, cloud infrastructure and third-party AI models. This creates challenges in tracing provenance, maintaining security controls as well as establishing clear lines of accountability across organisations. These conditions highlight the importance of establishing a shared language and practical guidance to help organisations assess risk continuously and consistently whilst implementing safeguards at the system level of their AI supply chain.

Addressing these challenges requires a proactive approach that builds on existing cyber security best practices while adapting them to reflect the layered and modular nature of AI systems. Strengthening resilience will depend not only on technical safeguards but also on structural enablers. This includes incentivising secure behaviour across the supply chain, building foundational knowledge within and across organisations and consolidating existing guidance into actionable resources tailored for operational needs. Ultimately, the broader goal is to equip both public and private sector with the tools, language and frameworks necessary to anticipate AI-specific risks beyond the surface level to address the foundational infrastructure on which we build 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.

CONTENTS

THE INVISIBLE STRUCTURE OF AI

Over the past five years, AI has transitioned from a niche capability to a critical enabler of economic productivity[1]. Corporations are racing to adopt it whilst adversaries aim to exploit it. However, modern AI systems do not stand alone. They are assembled, trained, and deployed through a latticework of global components and actors. Mimicking all global supply chains, the AI supply chain is only as strong as its weakest and least visible link. As a result, alongside ongoing discussions on "AI alignment", where goal orientation and aligning intent of AI systems is at the forefront, there is a growing recognition of the need to address the security of the AI supply chain.

A supply chain is the full ecosystem of constituent elements – businesses, suppliers, services, people, processes, data and infrastructure – that enable a product or capability to move from manufacture origin to the end user[2]. In the context of AI, this includes not only tangible assets such as hardware and compute infrastructure but also dynamic entities, such as data flows and deployment environments. As organisations become increasingly digitally interconnected and interdependent within supply chain ecosystems, the AI supply chain introduces a new dimension of systemic risk. This has led to a growing recognition of supply chain cyber security within business resilience, including supply chain risk management[3,4]. Risks within supply chains can take many forms including, economic, environmental, political and ethical disruptions that have the potential to interfere with the reliable flow of services and goods across a network[5]. In turn, cyber threats, like other types of disruption, are dynamic in nature and can propagate across interconnected systems often resulting in "cascading or ripple effects" that extend beyond the initial point of impact[6]. Understanding these dynamics enables organisations to identify where vulnerabilities exist, how they may be exploited and what measures can limit their effects.

The World Economic Forum's Global Cybersecurity Outlook 2025 highlights supply chain interdependencies as one of the six key factors contributing to the complexity of today's cyber security landscape[7]. It notes that growing reliance on interconnected supply chains

[1] Hu et al., Understanding Large Language Model Supply Chain: Structure, Domain, and Vulnerabilities. *arXiv*. April 2025. (Accessed September 2025).
[2] Building a Supply Chain Ecosystem for Competitive Advantage, GEP, January 2023 (accessed September 2025).
[3] Rohland, et al., The Cyber Resilience Compass: Journeys Towards Resilience. *World Economic Forum*. April 2025.
[4] Agrafiotis, et al., Unpacking Cyber Resilience. *World Economic Forum*. November 2024.
[5] The Use of AI to Tackle Supply Chain Risk. *CWSI*. 2025. (accessed March 2025).
[6] Ghadge et al., Managing cyber risk in supply chains. *Supply Chain Management: An International Journal*. 2020.
[7] Joshi, et al., Global Cybersecurity Outlook 2025. *World Economic Forum*. January 2025.

introduces increasing uncertainty and risks across sectors[8] (Tuteja, 2025). While many of these risks are shared between digital systems, the AI supply chain also presents distinct challenges. Literature points to unique vulnerabilities that differ in nature from those in traditional software chains, such as data poisoning, model obfuscation, indirect prompt injection and operational differences associated with data management[9] (ETSI, 2025). Today, many AI systems are vulnerable because the supply chains that support them are often opaque. Recent incidents, including compromised training libraries reveal how AI systems are exposed to manipulations and silent failures[10]. These vulnerabilities compose strategic risks, capable of jeopardising public trust in AI systems, regulatory compliance and national security.

Similar to well-known cyber security supply chain attacks, we envisage future equivalent AI attacks. Notable examples of attacks at various layers of the AI supply chain include:

Malware at the chip manufacturer TSMC (2018) [11] **-** Taiwan Semiconductor Manufacturing Company (TSMC), a leading producer of AI chips, was targeted through a ransomware attack that exploited vulnerabilities in a third-party IT services provider through software updates. This highlights the vulnerability of upstream hardware dependencies.

SolarWinds supply chain attack (2020)[12] - Hackers broke into the systems of SolarWinds, a company whose software is used by many governments and corporations. They inserted malicious code into a software update, silently spreading access to thousands of networks.

PyTorch incident (2022) [13] - Attackers uploaded a malicious Python package named "torchtriton" to the PyPI repository, mimicking a legitimate PyTorch component and stealing environment data from users.

[8] Tuteja, A., 5 risk factors from supply chain interdependencies in a complex cybersecurity landscape. *World Economic Forum*. January 2025. (accessed September 2025).
[9] ETSI, Securing Artificial Intelligence (SAI): Baseline Cyber Security Requirements for AI Models and Systems. ETSI Technical Specification. April 2025.
[10] PyTorch Foundation, Compromised PyTorch-nightly dependency chain between December 25th and December 30th, 2022, December 2022 (accessed September 2025).
[11] Banker, S., The Wolrd's Most Vulnerable Supply Chain Impacts All Supply Chains, *Forbes*, February 2023. (accessed September 2025).
[12] SolarWinds, NCSC Annual Review 2021, November 2021. (accessed September 2025).
[13] PyTorch.

LLaMa model leak (2023)[14] **-** Meta's LLaMA foundation model was leaked onto public platforms, including GitHub and Hugging Face, raising concerns regarding uncontrolled and open-source dissemination of AI models.

MCP OAuth vulnerability (2025)[15] **-** a widely used OAuth proxy package was compromised and republished with malicious files, enabling remote code execution on developer environments across MLOps pipelines. This was one of the first end-to-end AI pipeline compromises, showing how trusted tools can become risks, especially when components are published as open-source.

These incidents illustrate that each AI supply chain layer introduces distinct risks and vulnerabilities that can silently propagate to other systems and sectors if left unchecked. Addressing these risks requires a shared understanding of what AI systems are made of, who builds them and how they can be made more resilient. This can be achieved by establishing a common language and framework to describe the AI supply chain's components, interdependencies and potential points of failures, and can support a coordinated risk assessment and build resilience across both public and private sectors.

WHITE PAPER SCOPE AND MISSION

This white paper proposes a working taxonomy of the AI supply chain, maps attack surfaces across layers and interdependencies, and draws on sectoral insights and real-world cases to address risks and mitigations to support a more accountable AI adoption ecosystem. The findings and perspectives of this paper are primarily informed by discussions held during a workshop with a diverse stakeholder group. Supporting data was collected from a pre-workshop survey completed by 18 participants from 13 organisations. The whitepaper is not intended as a comprehensive catalogue of all AI supply chain risks and mitigations. Workshop participants included experts from across government, industry and startups, from a range of industries including Automotive and Transport, Finance, Cybersecurity, Technology, Logistics and National Security and Defence (refer to Appendix).

[14] Vincent, J., Meta's powerful AI language model has leaked online – what happens now? *The Verge*, March 2023. (accessed September 2025).
[15] Raina, A.S., MCP Horror Stories: The Supply Chain Attack, *Docker*, August 2025. (accessed September 2025).

TAXONOMY OF THE AI SUPPLY CHAIN

While often reduced to models and training datasets, the AI supply chain extends far beyond these elements, encompassing the software, hardware, infrastructure, and human inputs that make modern AI possible. Thinking of the supply chain in this way highlights not only its complexity, but also its vulnerability: as a vast system of interconnected components that together expand the attack surface of AI.

78% of workshop participants indicated their organisation had limited or partial visibility of their AI supply chain, highlighting the need for transparency and a common language shared between sectors. To effectively understand and manage this landscape, it is critical to establish a taxonomy of the AI supply chain. Such a taxonomy provides clarity on the components involved and the common language used to describe them, enabling more precise discussion of risks, dependencies and security challenges across the ecosystem. However, AI supply chains are not all uniform and must be examined in the context of their deployed environment. In addition, with increased adoption of single agent and multi-agent systems, the complexity of components involved in an AI supply chain is likely to change with the addition of different types of memory (episodic, short term or long term) or tool use, for example.

NO ONE-SIZE-FITS-ALL AI SUPPLY CHAIN

Organisations adopt AI through a wide range of models, including in-house development, use of open-source components, deployment of pre-trained models or reliance on APIs and hybrid approaches across on-premise and cloud infrastructure. Each of these deployments introduces distinct trade-offs in terms of visibility, control and vulnerability. For example, organisations building AI systems internally may have greater oversight of data pipelines and model logic, while those integrating external APIs must navigate opaque model behaviour and service provider dependencies. These variations shape both the technical risk and the accountability or governance structures surrounding the AI supply chain, leading to varying vulnerabilities and available mitigations[16,17].

[16] Cobbe, Veale and Singh, Understanding accountability in algorithmic supply chains. *FaccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability and Transparency*. June 2023.
[17] Widder and Nafus, Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data and Society*. June 2023.

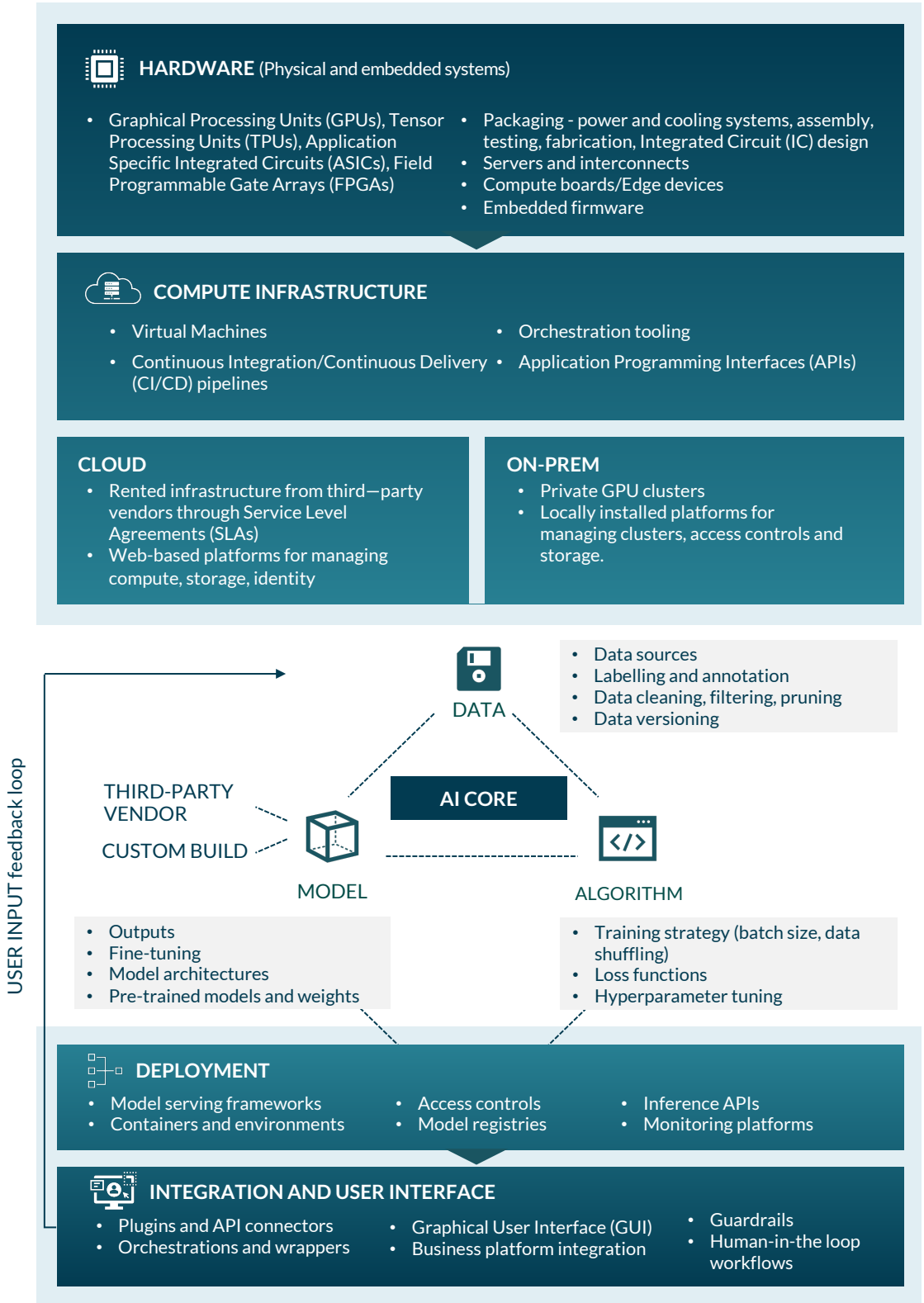


Figure 1. Layered taxonomy of the AI supply chain and its sub-components.

Several recent studies have advanced efforts to define and structure the AI supply chain. Work led by the Bank for International Settlements has focused on a macro-level taxonomy and systemic exposures related to infrastructure concentration[18]. Technical communities, such as the Coalition for Secure AI (CoSAI), have proposed risk-and-control based frameworks that map vulnerabilities across AI development and deployment stages[19]. Concurrently, regulatory research has taken an alternative approach by exploring the implications of oversight, while European Union–centred analyses have examined how the AI value chains intersect with legal definitions of providers and deployers under the EU AI Act[20]. These contributions have informed this work which builds on shared principles while proposing a taxonomy intended to support cross-sector understanding.

In practice, many organisations engage with the supply chain indirectly through third-party AI services, such as model Application Programming Interfaces (APIs) or pre-built components. These services often abstract away multiple underlying layers such as data training infrastructure or deployment environments which remain invisible to the end user organisation. While this simplifies adoption it also creates blind spots. Adopters may have little or no visibility into how models are built, what data they rely on and how they are maintained or updated. As a result, assumptions about control and accountability within the AI supply chain may no longer apply. Acknowledging that risk can be embedded across bundled, outsourced services is important for reassessing governance expectations, procurement strategies and the scope of internal security measures by adopting teams. Workshop participant organisations indicated that 28% had changed vendors due to trust or compliance challenges highlighting that services that may offer organisations simplified integration may in fact come at the cost of visibility and control across the AI supply chain.

Figure 1 presents a layered breakdown of the AI supply chain, outlining its five main components along with example sub-items within each.

HARDWARE

The hardware layer is the physical and embedded systems required to operate an AI system, including the general purpose computational components such as Graphical Processing Units (GPUs) that provide the processing power for an end-to-end AI system.

[18] Gambacorta and Shreeti, The AI Supply Chain. *Bank for International Settlements*. March 2025.
[19] Coalition for Secure AI (CoSAI), Establish Risks and Controls for the AI Supply Chain, V 1.0. *OASIS Open Project* (accessed September 2025).
[20] Engler and Renda, Reconciling the AI Value Chain with the EU's Artificial Intelligence Act. *CEPS In-depth Analysis*. September 2022 (accessed September 2025).

COMPUTE INFRASTRUCTURE

The compute infrastructure layer covers the platforms and systems that enable AI development and operations at scale. This includes how systems are virtualised on cloud infrastructure, how other systems connect to the compute layer via APIs, Continuous Integration / Development (CI/CD) pipelines and the hosting environment. The organisation's level of control over this layer should be considered, as well as whether these systems are deployed on-premises or in a private or third-party cloud infrastructure.

AI CORE

The AI core represents traditional thinking around AI and model operationalisation. This layer includes the AI model, training data and the algorithms that define how the model learns and operates.

1. ***DATA***

   Data represents both the raw inputs and the curated datasets used for training and evaluation. It can include a variety of expressions of data, such as structured or unstructured datasets, synthetic data and real-time streams collated from sensors and user interactions. Data shapes how the model interprets inputs to directly generate outputs, therefore the quality and composition of the data are central to the model's behaviour.

2. ***ALGORITHM***

   Algorithms define how a model is trained, optimised and evaluated. It defines approaches to training, such as classical statistical methods to deep learning architectures, which actively shape performance and enable explainability and robustness of AI systems.

3. ***MODEL***

   Models are the underlying technical architecture that responds to both data and algorithms used to create and respond to knowledge. They may take the form of large-scale foundational models, domain-specific models or compressed models for lightweight systems. The intellectual property of a model is extremely valuable and therefore a critical risk surface for model theft, tampering or reverse engineering.

DEPLOYMENT

The deployment layer represents the processes and systems required to move a prototype AI system into a live environment, and the management of that system throughout its lifetime. It includes how the organisation manages all components of the compute layer,

from cloud environments to on-premises hardware as well as the management infrastructure governing the end-to-end AI system.

INTEGRATION & USER INTERFACES

This layer captures how AI systems connect with end users, external applications and broader organisational systems. This includes APIs, Software Development Kits and user-facing applications that enable model functionality and support the user input feedback loop.

GREY ZONES

In mapping infrastructure to the AI supply chain taxonomy, several "grey zones" emerged where accountability and control are either fragmented or unclear (Figure 2). These often arise at the intersection of technical systems and deployment environments. As a result, "grey zones" concern both technical and organisational boundaries as they represent convergence points for risk propagation or failure that are challenging to trace.

CROSS-CUTTING DEPENDENCIES

The AI supply chain is shaped by a set of cross-cutting dependencies that influence its components and can introduce risks. As the body of knowledge surrounding AI supply chains is still nascent and rapidly evolving, it is essential to adapt and refine models in response to emerging insights.

TRADITIONAL SUPPLY CHAIN MONITORING

Mapping the AI supply chain is a sub-item of wider supply chain risk management. Therefore, incorporating the principles of supply chain security seen within the National Cyber Security Centre's (NCSCs) supply chain guidance remains critical[21].

CUSTOMER & STAKEHOLDER ALIGNMENT

AI systems must proactively address the needs and expectations of customers, end-users, and broader stakeholders. Ensuring alignment across these groups in areas such as data collection, model design, and deployment practices is essential to minimise risks and prevent erosion of trust that can result from misalignment.

[21] National Cyber Security Centre (NCSC), Supply Chain Security Guidance, January 2018. (accessed September 2025).

REGULATORY LANDSCAPE

The regulatory environment in which an AI system operates is a critical factor shaping the AI supply chain. Emerging frameworks such as the EU AI Act[22], ISO 42001[23] and sector specific regulations govern key elements including data collection, model evaluation, and deployment standards. Compliance gaps in these areas can expose organisations to both operational and reputational risks.

[22] European Union Artificial Intelligence Act (EU AI Act), July 2024. (accessed September 2025).
[23] International Standard, ISO/IEC 42001:2023. Information Technology – Artificial Intelligence – Management System, December 2023. (accessed September 2025).

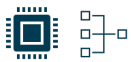
GREY ZONE	MAPPING	DESCRIPTION
Edge Computing and End-of-Life devices		Edge computing enables local AI processing on edge devices such as IoT sensors. This reduces latency however edge devices often lack patching, secure firmware or real-time monitoring due to limited storage, compute or power capacity. End-of-Life (EOL) devices amplify this challenge as unsupported, unpatched hardware remain active in critical systems.
DevOps tooling		CI/CD systems commonly pull dependencies, scripts and packages from public registries. Several components operate in the background where updates are often triggered automatically. These multi-stage pipelines may also be maintained by different teams or vendors. The complexity and lack of visibility make DevOps tooling a hidden exposure point.
Federated AI and Distributed Systems		Federated AI systems perform model training across decentralised devices , to preserve data privacy by storing data locally. However, traditional perimeter monitoring systems are unfit to monitor federated systems. As a result, this limits visibility into data quality and potential adversarial activity.
Open-Source dependencies		Open-source frameworks e.g. PyTorch, packages and tools are widely used in the developer community and frequently updated. However, they lack formal oversight. They can expose a system to malicious code injections through hidden or indirect dependencies that are difficult to detect and can silently propagate through the supply chain.
Data and annotation services		Outsourcing data collection and annotation to third-party vendors, contractors, or crowdsourcing platforms can obscure the source and quality of training data. Once used in model development, errors or inconsistencies become difficult to trace, with direct implications for downstream model behaviour.
Synthetic Data		Synthetic data generated by Large Language Models (LLMs) or Generative Adversarial Networks (GANs) may lack quality standards, introduce hallucinations, artefacts and poisoned data. Th lack of ground truth creates uncertainty and compliance concerns downstream in the supply chain.
Pre-trained Models and On-device inference		Pre-trained models can be sourced from third-party vendors or open repositories. This offers limited transparency into training data or biases. Furthermore, on-device inference adds risks such as model exfiltration. In both cases, organisations inherit model behaviour without full understanding or control.
Model-as-a-Service APIs		Model-as-a-Service offers AI models as cloud-hosted APIs. Whilst this enables rapid AI adoption, it results in the organisation having limited oversight of the model's training data, output behaviour and background infrastructure.
Agent ecosystems		Agent ecosystems e.g. LangChain extend the capabilities of foundational model capabilities through third-party plugins to allow for dynamic behaviour. This may introduce unverified components downstream vulnerabilities. As agent complexity grows, safety assurance lags as oversight struggles to scale.
Inter-organisational collaboration	Cross-cutting	Collaboration between public and private organisations often involves shared data, models or infrastructure without collaborative risk governance. Contractual obligations and unclear accountability render it difficult to manage compliance and incident response across organisational boundaries.
Shadow IT		Unofficial use of AI tools and models, such as LLMs, created blind spots for security teams, compliance and threat monitoring. Shadow IT remains an unmanaged component which may enter critical workflows.

Figure 2. Grey zones in the AI supply chain and the supply chain mapping to a layered taxonomy.

VULNERABILITY LANDSCAPE

Propagation of the vulnerabilities on the AI supply chain is a systemic challenge cutting across several dimensions of the supply chain, including the provision of the hardware resources, provenance and governance of the data, fine-tuning and Retrieval-Augmented Generation (RAG) resources, integration of and reliance on third-party systems and user interfaces. When the data, model and deployment systems are re-used without further validation, these vulnerabilities can cascade across deployments downstream (Figure 3). This list is not intended to be exhaustive but instead reflects the collective insights and shared understanding that emerged from the workshop discussion.

RISKS

Workshop participants highlighted emerging AI supply chain risks that may be neglected in current risk assessment practices across their organisation. By risks we refer to the possibility of negative consequences to an organisation, system, or society that result when vulnerabilities are exploited, controls fail or threats materialise.

Traditionally, the CIA triad (Confidentiality, Integrity, Availability) defines the pillars of information security. However, in the context of AI, additional properties — including explainability, fairness, robustness, reproducibility, and accountability — are equally important for ensuring systems are fit for purpose and safe. Insights from the workshop suggest that these properties offer a practical lens for assessing AI supply chain risks than component-level analysis alone. By taking a unified view, both isolated failures and the breakdown of these principles across interconnected supply chain layers lead to the highest risk categories. This perspective suggests a shift in how supply chain security is evaluated to focus on how both secure-by-design and sourcing decisions affect the overall reliability and responsible use of AI technologies by adopters.

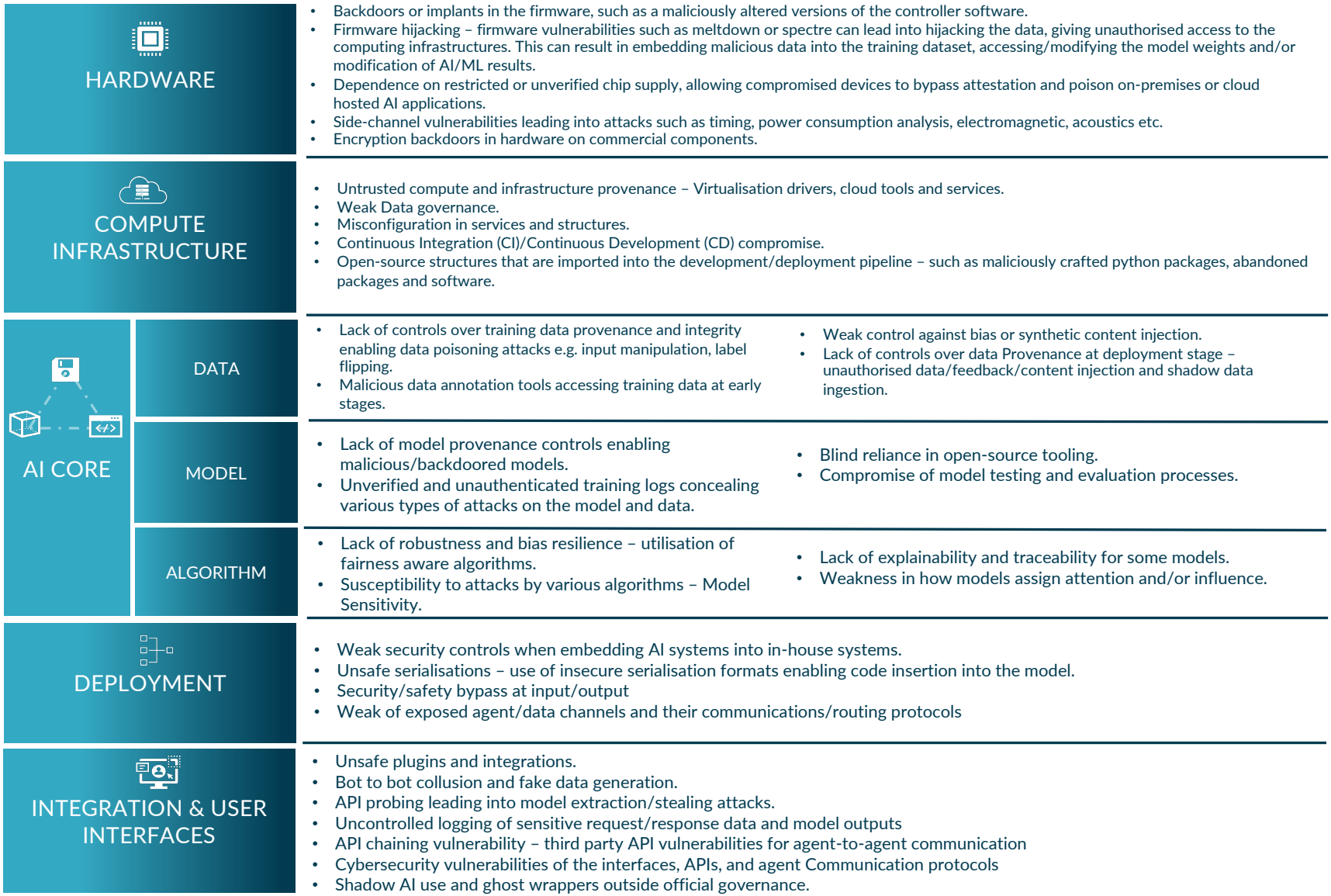


Figure 3. Vulnerability considerations mapped to the AI supply chain.

RISKS TO ORGANISATIONS

OPERATIONAL

- Service outages or disruptions
- Supply chain or system integration failures
- Infrastructure breakdown (dependency failures)

FINANCIAL AND LEGAL

- Liability and regulatory fines from non-compliance
- Legal or rights violations (IP, licensing, unlawful data use)

STRATEGIC

- Mission or safety-critical failure
- Model IP theft (loss of competitive advantage)

REPUTATIONAL

- Compromise of customer-facing / user mistrust
- Brand damage from unsafe or biased outcomes

RISKS TO SOCIETY AND INDIVIDUALS

PRIVACY: leakage of sensitive personal data.

FAIRNESS: discrimination in hiring, credit, healthcare, or access to services.

SAFETY: reliance on hallucinated, corrupted, or unsafe outputs in critical contexts.

TRUST: misinformation or inability to contest/understand automated decisions.

ETHICAL: erosion of accountability, normalisation of bias.

SOCIETAL: : large-scale misinformation, reduced public trust in AI systems.

GEOPOLITICAL: Balkanisation of AI ecosystems, restricted market access

CHALLENGES OF SECURING SUPPLY CHAIN DEPENDENCIES

Organisations face a few challenges specific to the AI supply chain, many of which centre on limited visibility and assurance. These include a lack of transparency into security practices of third-party service providers, unclear accountability across the supply chain and difficulty in obtaining assurance due to lack of metrics and standards for assured service provision. In addition, many systems rely on unvetted open-source components, including data, model architectures and supporting libraries, without provenance tracking.

At the organisational level, potential challenges include rapid scaling of AI initiatives without appropriate controls, the ease of spinning up uncontrolled Proofs-of-Concept and the general lack of AI security governance or strategy within organisations. Several leaders report a lack of awareness of AI-specific security risks and teams often struggle to assess the threat landscape. Shadow AI and shadow data practices, where models are developed or used without oversight, are increasingly common alongside immature resilience practices and a tendency for employees to bypass security policies in favour of speed or innovation.

MITIGATIONS

Typical steps to reduce the overall risk of utilising a system involve education, technical controls, governance and incentives to reduce exposure.

For AI supply chain risks, this may look like:

- Securing procured AI components such as hardware, model architecture, datasets (procurement controls, testing).
- Securing a trained model that has been procured or developed and trained in-house (audit controls).
- Ensuring security of AI services used (controls for service usage e.g., security requirements in contracts).
- Securing training and the fine-tuning process (training time and training dataset controls).
- Securing an operational model by applying controls and processes to protect against, detect, and recover from threats.

To support this, there is a need for:

ASSURANCE FRAMEWORKS AND STANDARDS

To build trust and resilience in AI systems there is a growing need for robust assurance frameworks that address both products and services within the AI supply chain. For procured components, such as hardware, model architectures and datasets this includes the implementation of procurement controls and testing protocols to validate security prior to deployment. On the other hand, for models developed or acquired by a third-party, mitigations should involve rigorous testing and evaluation procedures to detect in-built vulnerabilities, biases or performance degradation. However, the assurance of AI services is an additional consideration, where contractual agreements for providers require clear security controls for use of services. These efforts should be underpinned by widely accepted standards and best-practice guidelines, such as International Organization for Standardization (ISO), National Institute of Standards and Technology (NIST) and industry bodies, that support the application of technical controls to protect, detect and recover AI systems.

EDUCATION

Educating stakeholders ensures they understand AI risks and opportunities. For C-suite, this means understanding how to balance risk vs. reward in adopting AI [cite WEF report]. For technical teams, it means the ability to use tools and processes to support secure AI adoption such as model cards and provenance tracking. There is also a need to build an understanding of how to communicate AI supply-chain security risks within the organisation.

The educational body of knowledge for AI cybersecurity risk is developing but not yet clearly defined. There is a need to build resources to help SMEs, corporates and governments foster AI security risk awareness and communication. This means building understanding of how to aim the right information at the right roles without creating information deluge.

CASE STUDIES

The following case studies demonstrate how the AI supply chain taxonomy can be applied to analyse real-world security incidents. By mapping each case to specific layers, tactics and vulnerabilities, the taxonomy helps identify threat entry points, risk propagation across layers and possible mitigations. These types of examinations can turn isolated events into insights for resilience planning across organisations.

Case Study 1: Backdoors in the Supply Chain

Backdoor Data Poisoning attacks involve the fabrication or modification of training, fine-tuning or external-inference time data to embed triggers or backdoors during the training or fine-tuning phase. These backdoors in the model can then be triggered at inference time, to cause manipulation of output data (e.g., to cause misclassification of certain inputs). This creates a risk of operational, financial or reputational harms arising from a loss of model-output integrity.

Parts of the taxonomy that should be considered for each attack vector are provided in the parentheses. Example attack vectors may include:

- TRAINING DATA INJECTION: Attackers embed the backdoor from the beginning/creation of the dataset. *(A. Data Component, B. Model Component).*
- FINE-TUNING WITH MALICIOUS DATASETS: Attackers weaponise the fine-tuning dataset for propagation of the backdoor. *(C. Model Component)*
- RAG POISONED DATASETS: Inheritance of the backdoor into the inference time (post-deployment) decisions. *(Deployment Component)*
- WEIGHT POISONING**:** A white-box attack – attacker has access to the weights of the model and manipulation is done at the low-level *(Model and Algorithm Component)*
- SHADOW DATA INGESTION: Feeding poisoned data into the deployed systems by malicious users on highly regulated architectures (Deployment, Integration and User Interface).

MAPPED WEAKNESSES, TACTICS, TECHNIQUES AND PROCEDURES (TTPs) AND INDUSTRY-WIDE DEFINITIONS:

MITRE CWE	CWE-349: Acceptance of Extraneous Untrusted Data with Trusted Data CWE-506: Embedded Malicious Code
ATLAS	AML.T0018: Manipulate AI Model AMLT0018.001: Manipulate AI Model: Modify AI Model Architecture AML.T0020: Poison Training Datasets
OWASP	LM04:2025 Data and Model Poisoning

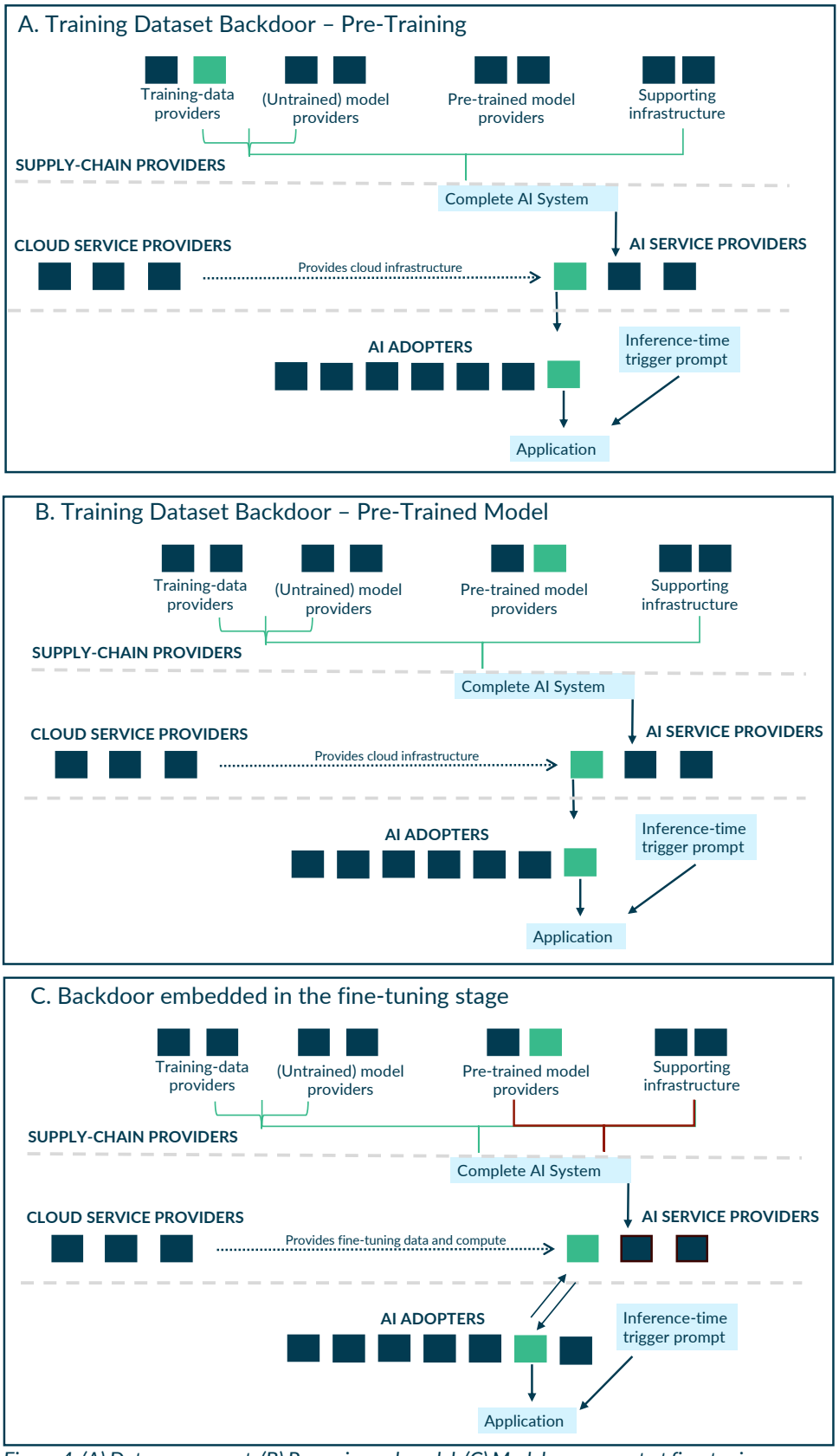


Figure 4. (A) Data component. (B) Pre-poisoned model. (C) Model component at fine-tuning.

THREAT MODEL

The attacker will first deploy the vulnerability through access to datasets and then exploit the AI system using either a prompt injection through a legitimate account or using agents trained on/utilising the foundational model provided and triggered externally. This triggering mechanism can be defined as a specific input - can be random keywords or can be part of a datasets (role-play embedded in the AI system or can be a hidden parameter that can be obfuscated in the weights of the model. In Figure 4a, the vulnerability originates at the training data provider and extends all the way to the AI-embedded application. On the other hand, Figure 4b shows a model trained with poisoned data propagating downstream to the application. The attackers published a poisoned model and uploaded it into the model repositories of legitimate services. Finally, Figure 4c represents the case of a model poisoned during the service provision stage, where fine-tuning data or feedback is manipulated with maliciously crafted datasets. At this stage, the model may already have passed robustness and security/safety checks.

RISKS AND IMPACT: Loss of integrity of outputs: targeted misclassification, degradation of the system key performance metrics, degraded robustness.

MITIGATIONS: AIBOMs, which provide a detailed view of the materials that make up an AI system help ensure transparency in the supply chain. Known triggers for backdoors can be blocked and monitored through guardrail implementations. Continuous monitoring, anomaly checks, and robustness tests confirm resilience throughout the pipeline.

Case Study 2: Model Poisoning

The model's architecture is replaced with a maliciously organised architecture instead of fabricating the training data. This type of attack can unveil itself in attack vectors in the shape and format of:

- Architectural Backdoors: backdoors that comes with the poisoned model (Model Component).
- Federated Learning: Model poisoning via providing model updates. *(Integration and User Interfaces – Figure 5).*
- Imitating a genuine ML/AI model on publicly hosted model hubs such as HuggingFace *(Model Component - Figure 4b).*

MAPPED WEAKNESSES, TTPS AND INDUSTRY-WIDE DEFINITIONS:

MITRE CWE	CWE-502 Deserialization of Untrusted Data CWE-349 Acceptance of Extraneous Untrusted Data with Trusted Data
ATLAS	AML.T0010: ML Supply Chain Compromise AMLT0058: Publish Poisoned Models
OWASP	ML10:2023 Model Poisoning

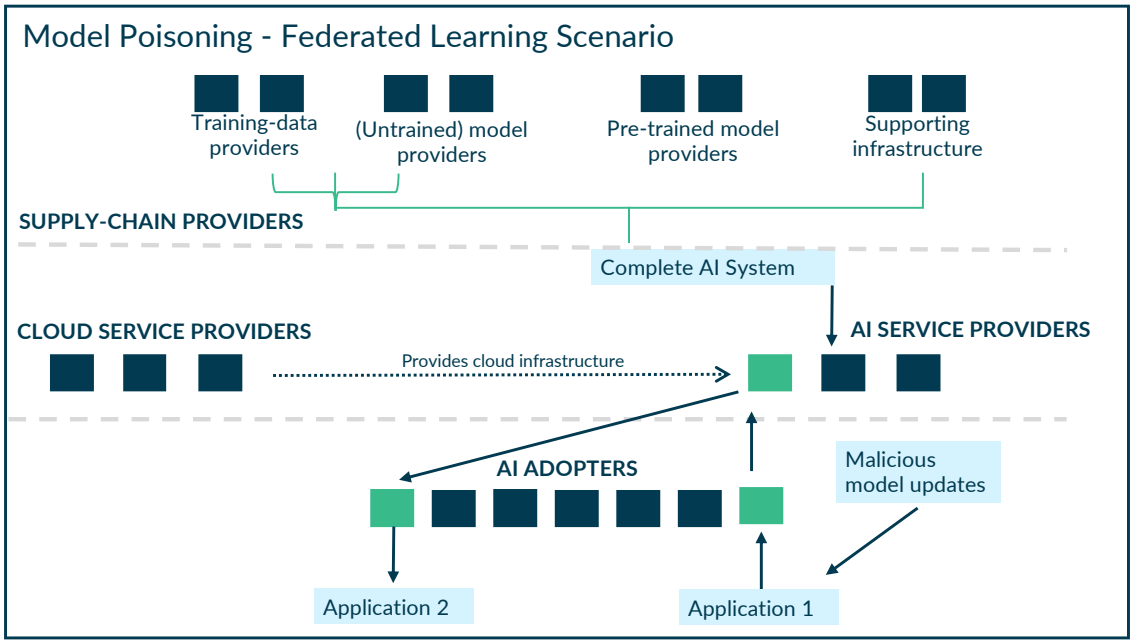


Figure 5. Case study 2 on model poisoning.

THREAT MODEL

The attacker can build a malicious model from scratch or fine-tunes one of these models with a poisoned fine-tuning dataset and uploads their model to one of the model hosting applications. The poisoned model can make incorrect predictions, degrade the overall performance of the system, or be trained specifically to leak sensitive information.

RISKS AND IMPACT: Loss of integrity of outputs: hidden backdoors, takeover of federated learning, supply-chain compromise.

MITIGATIONS: AIBOMs ensure transparency around models, datasets, and dependencies, while continuous monitoring and auditing help detect and resolve supply chain issues in real time. In addition, cryptographic methods can be used to verify the authenticity of AI models.

Case Study 3: Third-party API Compromise

In this attack scenario, the vulnerability lies not within the model or data, but in the interfaces that connect AI systems to external services. This type of compromise can manifest through several attack vectors (Figure 6):

- **Malicious API Implementations**: Attackers publish seemingly legitimate API wrappers that include hidden data exfiltration or logic manipulation routines through public code repositories such as GitHub (Integration and User Interface).
- **Man-in-the-Middle API Hijacking**: Intercepting API calls between AI systems and external services to alter responses or inject adversarial inputs (Integration and User Interface).
- **Over-privileged Integrations**: APIs granted excessive permissions can be abused to access or modify model behaviour beyond intended scope (Integration and User Interface).
- **Shadow APIs**: Unmonitored or undocumented APIs within the supply chain can serve as hidden entry points for attackers (Integration and User Interface).

These compromises often go undetected due to the trust placed in widely used APIs and the complexity of monitoring real-time interactions. As AI systems increasingly rely on external services for data enrichment, inference, and deployment, securing API interfaces becomes a critical priority.

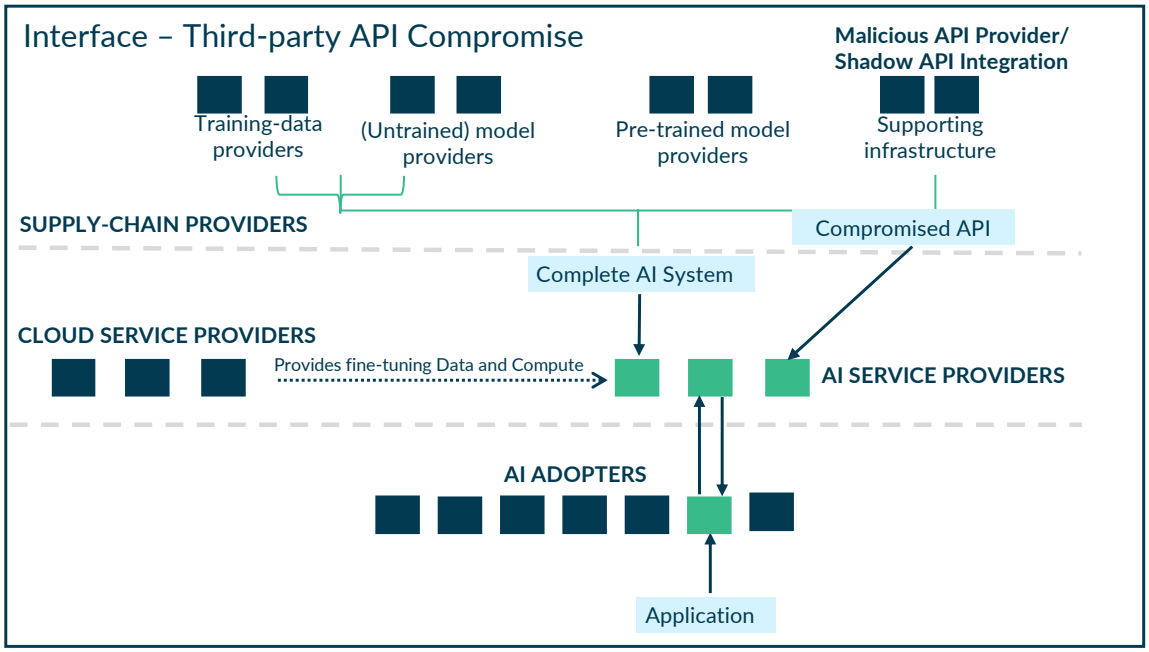


Figure 6. Case study 3 - Third-party API Compromise

MAPPED WEAKNESSES, TTPS AND INDUSTRY-WIDE DEFINITIONS:

MITRE CWE	CWE-648: Incorrect Use of Privileged
ATLAS	AML.T0040: AI Model Inference API Access
OWASP	ML10:2023 Model Poisoning

THREAT MODEL

Malicious actors exploit third-party APIs to inject harmful payloads, intercept sensitive data, or manipulate model outputs.

RISKS AND IMPACT: Data leakage, loss of integrity, loss of availability, over-privileged applications through APIs

MITIGATIONS: For tracking API and third-party activity in systems, events, metrics, and traces from third-party elements are logged to detect unusual behaviour. To prevent unauthorised access and keep endpoints secure, authentication for APIs is enforced, and to reduce the risk of unsanctioned wrappers or shadow APIs, source code controls are applied.

KEY TAKEAWAYS

As AI systems become increasingly embedded in CNI and commercial products, ensuring the resilience and security of the AI supply chain is foundational to maintain secure and robust future systems. Findings from this work suggest that the AI supply chain consists of layered and interdependent components, where risks may emerge both on individual elements, such as data sources, as well as the interactions between them. Areas of limited visibility or unclear responsibility, such as the integration of third-party services and open-source tools on customer-facing platforms, present ongoing challenges for organisation in terms of assurance and oversight. Stakeholder input reflected a broad view that existing cyber security and supply chain risk practices provide a useful foundation. Adapting these to the specific characteristics of AI systems, such as dynamic model behaviour, will be unlock improved resilience in practices for all sectors in a practical and scalable way.

NEXT STEPS

- EXISTING CYBERSECURITY: Build on existing cybersecurity, supply chain and risk governance frameworks (e.g. ISO, NIST) and adapt them to the AI context.
- INCENTIVISE SECURITY ACROSS THE SUPPLY CHAIN: Create meaningful drivers, including regulatory, financial and reputational, for all to adopt and adhere to secure practices.
- DEVELOP FOUNDATIONAL KNOWLEDGE: Invest in training and educational programmes to upskill users, policymakers and decision makers within industry in AI-specific security and risk management.
- CREATE AWARENESS OF BODIES OF KNOWLEDGE: Ensure that best practice guidance, standards and resources are well-communicated and consolidated to provide a common language and avoid fragmentation or information overload.
- ACCOUNTABILITY AND REGULATION: There is a need to define accountability for security across the components of the AI supply chain. This may include the roles of insurance and regulation (AI-specific, sectoral and data protection).

Strengthening the AI supply chain will not only require technical solutions but it will rely on improving transparency, clarifying responsibilities and fostering collaboration across sectors. At present, there is limited shared responsibility of risk between service providers and users of AI systems, even as global interdependencies continue to shape how AI systems evolve e.g. agentic AI. Addressing these challenges calls for a proactive approach that builds on established best practices while introducing adaptable mechanisms suited to the AI systems adopted by individual organisations. This whitepaper contributes to ongoing efforts to establish common frameworks and shared understanding in order to support transparency and accountability within the AI supply chain.

APPENDIX

METHODOLOGY

To support the development of this whitepaper primary and secondary research was carried out in the form of data collection as a pre-workshop survey, a workshop held in September 2025 and desk research, respectively.

The workshop survey captured in multiple choice questions the AI adoption or provider profile, sector, organisation size and awareness of the organisation's supply chain. It was completed by 18 participants from 13 organisations. The survey captured the following sectors: Automotive and Transport, Finance, Cyber security, Technology, Logistics, National Security and Defence (NS&D), Government and Academia.

- Government
- Cisco
- Surecloud
- Lloyd's Banking Group
- APH10
- Fortinet
- Canopy
- Microsoft
- Athenian Tech
- U.S. Embassy
- BAE Systems (Digital Intelligence)
- BAE Systems
- The Alan Turing Institute
- Global Cyber Security Capacity Centre (University of Oxford)

The workshop session brought together 17 participants from 10 organisations, including LASR partners and followed Chatham House Rules. Structured workshop activities aided the discussion on the AI supply chain taxonomy, risks and mitigations. Experts in the room ranged in experience and sectors including NS&D, Cyber Security, Technology, Government, Finance, Academia and Consulting services.

- Government
- Cisco
- Surecloud
- Lloyd's Banking Group
- BAE Systems
- Microsoft
- Plexal
- U.S. Embassy
- BAE Systems (Digital Intelligence)
- Global Cyber Security Capacity Centre (University of Oxford)

AUTHORS

PLEXAL

Plexal is the innovation and growth company helping to strengthen the UK's technology capabilities through collaboration. With teams in London, Manchester and Cheltenham, we work closely with government, industry, startups and academia to drive economic growth and reinforce national security. Founded by Delancey in 2017, Plexal delivers four core services: creating workspaces for innovators, running innovation programmes and consultancy, building regional tech clusters and helping SMEs scale strategically. By closing the gap between early-stage and established organisations – across the public and private sectors at home and abroad – we've impacted over 1,200 businesses, added £731m to the UK economy and helped create 9,400 jobs. Headquartered at Here East on the Queen Elizabeth Olympic Park, Plexal partners with organisations including the Department for Science, Innovation and Technology, the National Cyber Security Centre, Airbus, Amazon, Barclays Eagle Labs, Google Cloud, TfL, The Alan Turing Institute and the University of Oxford.

GLOBAL CYBER SECURITY CAPACITY CENTRE

The Global Cyber Security Capacity Centre (GCSCC) is a leading international centre for research on efficient and effective cybersecurity capacity-building, promoting an increase in the scale, pace, quality and impact of cybersecurity capacity-building initiatives across the world.

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 cyber security, develop novel capabilities and skills, accelerate research commercialisation, and foster international collaboration for the secure development and deployment of AI.

Plexal

14 East Bay Lane, The Press Centre, Here East, Queen Elizabeth Olympic Park, London, E20 3BS

plexal.com

+44 (0) 203 909 7763

connect@plexal.com

Research conducted by: