Australian Government
Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC Australian Cyber Security Centre

Artificial intelligence and machine learning

Supply chain risks and mitigations

First published: October 2025
Last updated: October 2025

Table of contents

Introduction

Artificial intelligence (AI) and machine learning (ML) systems allow organisations to improve their efficiency in many areas. These systems can help inform decisions, streamline processes and improve customer experience. For an explanation of the AI-related terminology used in this guidance, refer to the Australian Signals Directorate's (ASD) publication, *Convoluted layers: An artificial intelligence primer*.

Adopting AI or ML systems also brings unique supply chain risks, which can threaten the cyber security of an organisation if not securely managed. The use of pre-trained open-source models and datasets from public websites can make these risks more pronounced.

This guidance is intended for organisations and staff that deploy or develop AI or ML systems and components. This could range from entirely outsourcing an AI system where an organisation only provides the training data, to in-house AI development. This guidance aims to:

- highlight the importance of AI and ML supply chain security
- address key risks and mitigations that should be considered when developing or procuring an AI system.

While this guidance provides some mitigation descriptions, cyber security staff should still consult detailed technical guidance and frameworks for implementation. To help with this, the risks highlighted below are mapped to the National Institute of Standards and Technology's (NIST) Adversarial Machine Learning (AML) taxonomy where applicable. Additionally, refer to the *More information* section at the end of this guidance.

This guidance focuses on aspects unique to the AI and ML supply chain, though general cyber supply chain risk management still applies. For more information, refer to ASD's *Cyber supply chain risk management* publication.

Overview

The AI and ML supply chain is complex. Vendors and service providers often need to source or manage various components. This includes the underlying AI training data, model, software, infrastructure, hardware, and third-party services.

Each component may contain vulnerabilities and risks that a malicious actor could exploit to compromise confidentiality, availability and integrity of the system. Depending on the component, these exploits could expose training data, degrade AI or ML system functionality, or execute malicious code.

Supply chain risk management concepts for AI and machine learning

Organisations should assess their AI and ML supply chain as part of their broader cyber security risk management strategy. Existing cyber security considerations that would apply to any new software, system or dataset still apply to AI and ML components. For more information on general cyber security management, refer to ASD's *Information security manual*.

Integrating AI and machine learning products and services

Effective risk management requires full visibility of AI and ML systems and their supply chain. Organisations should:

- identify suppliers, manufacturers, distributors, retailers and subcontractors
- seek information on any AI or ML features that are added to existing systems and consider adjusting cyber security risk management for the affected systems.
- consider how new functionality from AI or ML products could affect your cyber security risk management.

Working with third parties

When working with AI and ML vendors and suppliers, discuss cyber security considerations with them at an early stage, and understand shared responsibilities. Organisations should:

- prioritise supply chain components or vendors that are responsible for critical cyber security functions
- consider cyber security management for implementation and throughout the product's life cycle
- include cyber security requirements in contractual agreements and outline shared responsibility models.

Case study: Data leaks from third parties

In 2025 alone, there have been several high-profile data leakages due to compromise of third-party services. A recent survey found that around 65% of organisations have been impacted by an AI-related data leak, with another survey showing that 13% have reported breaches directly to their AI systems.

As AI use increases data throughput and aggregation, the impact of these kinds of leaks is set to increase as well.

Securing your organisation's data

Organisations should understand how AI and ML systems affect their data security by:

- considering the value of your data and what other information could be derived from it
- understanding what data you process, store and communicate to support AI or ML systems
- considering who has access to sensitive data, model inventory, data storage, infrastructure or hardware
- identifying and considering which AI and ML vendors can access your data and how they use it
- being aware of data that may be transmitted, processed or stored in an offshore location.

Communicating and training within your organisation

Incorporating AI and ML systems will likely require additional internal management for supply chain and staff training. Organisations should:

- establish clear communication channels for any AI and ML supply chain concerns
- educate all staff involved in AI or ML system development, deployment, use and maintenance about cyber security risks and best practices.

AI data

AI and ML systems typically use large datasets and knowledge bases that often need to be maintained for future retraining or new models. This data may be sensitive and is often stored in a single location so it can be efficiently retrieved when training the model. Training data may also be obtained through open sources, which comes with additional considerations regarding accuracy and security.

Conventional data security management should still be applied, including consideration of increased data aggregation. For more information, refer to ASD's joint *AI data security* publication.

Risks

This section covers some risks associated with data in the AI and ML supply chain.

Low-quality or biased data

Key question: Where do you source your AI training data? How do you ensure it is high quality and unbiased?

NIST AML Mappings: None (not an intentional attack)

Low-quality or biased data may be inaccurate or may not clearly represent the problem the AI or ML system aims to solve. Poor data collection or generation processes can lead to bias or inaccurate labelling, causing the resultant model to be harder to train and less effective overall. This can cause the resulting application to have more incorrect classifications or predictions, ultimately reducing robustness.

Data poisoning

Key question: How do you ensure your training data has not been maliciously tampered with?

NIST AML Mappings: NISTAML.012, .013, .021, .023, .024

Data poisoning involves malicious modifications to data that can affect the final AI or ML model and its outputs or cause unintended system behaviour. Data poisoning can be hard to detect due to the size and complexity of datasets, especially if data is already compromised at the source.

The effects of poisoned data on the resulting model can include:

- degraded performance and bias
- unintended or malicious responses to a specific data feature.

In other cases, the poisoned data can be used for indirect prompt injection. This is where the model mistakes part of the data as an instruction, which can lead to malicious behaviour.

Case study: Organisations affected by data poisoning

Joint reporting on *AI data security* confirmed that adversarial actors are already poisoning AI training data. Other research from 2025 found that around 25% of organisations had been a victim of AI data poisoning.

Training data exposure

Key question: How do you reduce the potential damage from exposed training data?

NIST AML Mappings: NISTAML.031, .032, .033, .034, .037, .038

Through manipulation or reverse engineering of AI and ML models, adversaries may be able to extract content or insights from data used to train the model. Even if the model can no longer access the original data, this information can be derived from the model's training memory. Attacks of this type include:

- model inversion (causing general insight into data)
- membership inference (if a particular sample is in the data)
- training data extraction.

These attacks could reveal sensitive information or otherwise be a privacy and confidentiality concern. While this is a risk primarily when using the model, organisations can apply several mitigations prior to model deployment to reduce risk.

Mitigations

Preventative measures are often more effective than detection for data risks. Mitigations should focus on ensuring the security, integrity and accuracy of the data from the outset.

Data collection and generation methods

Ideally, collection or generation of data should use a standardised method with supporting documentation. This can help with consistency and reduce the risk of bias, inaccuracy or other issues.

Data review and preprocessing

Organisations should review and preprocess data before using it in an AI or ML model. Data review should check that data is as expected and free of obvious errors. Preprocessing can help prevent bias by correcting issues such as missing features, duplicates or uneven sample distribution. These steps can help to ensure the data is of sufficient quality, though take care that they do not introduce bias themselves.

Data sanitisation

Sanitisation is a data-refining process that can include techniques such as filtering, sampling and normalisation. This process can reduce unwanted aspects in a dataset like noise, outliers and poisoned content.

Trusted sources

Source data from trusted and reputable providers where possible. Avoiding unknown sources can be a simple but effective way to reduce the risk of poisoned or maliciously modified data.

Data verification and provenance

Verification methods (such as checksums and digital signatures) are a way to check the integrity of digital data. Note that this only verifies integrity against the source and does not include any modifications made to the data prior to publishing.

More robust verification methods (such as lineage tracking and data provenance) are in development that will allow visibility of all changes on a dataset since its creation. With additional chain-of-custody information and tracking of changes, maliciously modified or poisoned data can be identified early. For an example of this in practice, refer to *Content credentials: Strengthening multimedia integrity in the generative AI era*.

Ensemble methods

Ensemble methods involve training multiple different datasets (or subsets) and models on the same problem to compare performance and consistency. This means that if a dataset or model is compromised, combinations using that component may have significantly different outputs to the others. These metrics can then prompt developers to further investigate, rectify or remove those components, reducing risks from data or models that are poisoned or of low quality. Orchestration tools can help automate this process and accurately assess system behaviour.

Training data obfuscation

Obfuscating training data removes identifiable or sensitive features while maintaining key aspects and patterns. Organisations should remove unnecessary or unimportant data and use anonymisation or differential privacy methods where possible. Where sensitive content is required, generating synthetic data based on the real content can reduce sensitivity while maintaining important features. While these methods do not prevent data extraction attacks, the obfuscation can reduce further damage.

Machine learning models

An ML model is the structure that transforms inputs into outputs based on patterns learned from training data. Attacks often involve manipulating the parameters within a trained model to cause an effect such as bias, degraded accuracy, or embedding of malicious code.

Risks

The following are some risks associated with the supply chain of ML models.

Model serialisation and similar attacks

Key question: How do you verify that a saved model is not compromised?

NIST AML Mappings: Part of NISTAML.05

Serialising, archiving and similar methods package ML models and associated components into portable and widely compatible formats. A model serialisation attack could occur when malicious code is added to the model package during the serialisation process. Other methods could add malicious code to a post-install script or preprocessing method, yielding similar effects with minimal user input. Malicious actors can then post the compromised file online to spread malware that activates when a user loads the contents. Malicious code added to associated components can cause similar effects.

Case study: Malicious ML models on sharing platform

Research from 2025 has shown that serialised model files containing malicious content continue to be uploaded to popular sharing platforms without being flagged by security checks. Due to the way many of these tools function, novel malware hiding methods can still go undetected.

Model poisoning

Key question: How do you ensure your model has not been maliciously modified?

NIST AML Mappings: NISTAML.011, .026, .051

Model poisoning refers to malicious modifications to an ML model. This can take various forms:

- Availability poisoning – degrades model performance.
- Targeted poisoning – adds bias or degrades specific aspects or categories of data.
- Backdoor poisoning – embeds a hidden trigger that cause the model to behave abnormally when activated, while performing normally otherwise.

Due to the complex structure of ML models, these modifications can be hard to detect before testing.

Model-based malware embedding (stegomalware)

Key question: How do you ensure a model is not hiding malware?

NIST AML Mappings: Part of NISTAML.05

Due to their relatively large size and complex structure, ML models can serve as hosts for malicious content that may avoid conventional malware detection. Malware can hide segments of code in the weights or metadata of a model that antivirus software may not typically check.

Loading or running a malicious model could allow reconstruction and execution of malicious code, even while maintaining the expected performance. Malicious actors can use this to stealthily run malware on a system or stage it for later use.

Evasion attacks

Key question: How do you design a model to be resilient to unusual inputs?

NIST AML Mappings: NISTAML.022, .025

Malicious actors can engineer inputs specifically designed to evade ML-based security mechanisms such as spam filters or intrusion detection systems. They often circumvent simple rule-based methods by changing their evasion techniques in response to new controls. This can cause a model to have unintended or malicious behaviour.

While some mitigations (such as monitoring and filtering inputs and outputs) can be implemented after deployment, steps should also be taken in the supply chain to improve general resilience to unusual inputs.

Mitigations

To mitigate risks in ML models, validate them at the point of retrieval and continuously monitor their performance.

Secure file formats

Secure AI and ML model formats help protect model content, saving and loading processes to avoid serialisation attacks and similar issues. Approaches include using non-executable weight formats or immutable graph formats, and applying safe loading flags such as weight-only. These methods rarely have significant downsides compared to conventional serialisation methods, so use them where possible.

Trusted sources

Like data, use models from trusted and reputable providers where possible. It is a simple yet effective way to reduce the risk of poisoned or maliciously modified models, but does not eliminate all risk. Treat all content sourced from the web as untrusted.

Model explainability and transparency

Explainable and transparent models help users understand how AI systems work and build trust in their outputs. These aspects can be enhanced through techniques such as model visualisation, feature importance analysis, and model simplification.

Improving explainability and trust also gives better insight into a model’s integrity and security, making it easier to identify issues like model poisoning. While this field is still evolving, it already benefits certain applications and is likely to become a key tool for AI systems as methods mature.

Model verification and provenance

Like AI data, use verification methods (such as checksums and digital signatures) to check the integrity of an ML model. For more information, refer to the section on *Data verification and provenance* under *AI data*.

Ensemble methods

Ensemble methods involve training multiple different datasets (or subsets) and models on the same problem to compare performance and consistency. For more information, refer to the *Ensemble methods* section under *AI data*.

Reproducible builds

A reproducible build refers to when developers release the components of a model rather than the complete and compiled version, like a software bill of materials (SBOM). By including build instructions with these components, users can create the desired final model and verify its integrity with a checksum or similar method.

By starting with the components instead of the full model, it is easier for users to interrogate the components, reducing the risk of poisoning or embedded malware. AI-specific extensions to existing SBOM, or purpose-built AI BOMs, are emerging and are expected to help with this approach in the future.

Testing and monitoring

Initial testing, along with ongoing input and output monitoring, helps ensure the model operates as intended. This may include periodically testing live models against expected or historical results, especially for models that use continuous learning.

Test and monitor model data to help identify issues such as model poisoning, model drift, or malicious use like evasion attacks.

Adversarial training

Adversarial training introduces an adversarial model during the training process to create or augment training data in unexpected ways. This data helps to test weaknesses in a model, which helps improve its resilience to unusual or malicious inputs.

By challenging the model, it strengthens its resistance to poisoned data and adversarial techniques after deployment. It can also improve a model’s overall performance.

Model refining

Similar to data, some techniques can be applied to a model to help reduce the risk of malicious modifications. Effective techniques include pruning, compression and fine-tuning. These methods alter or remove weights and neurons within a model, which can disrupt the structure of embedded malware, often at the cost of performance.

Applying these techniques before using models can help reduce some poisoning and malware-embedding attacks.

Security tools

As model risks become more widely known, both open-source and proprietary tools are emerging that can help mitigate some of them. Consider incorporating relevant ML-specific security tools into your cyber supply chain. For example, stegomalware scanners, similar to conventional malware detection tools, have proven effective against certain types of malware embedding methods.

A growing range of tools can now address many cyber security considerations for AI and ML noted in this guidance, though their effectiveness can be hard to verify.

AI software

AI and ML software includes any software that is used throughout the AI life cycle. This could be development frameworks, associated libraries, and other software that enables the system.

As with conventional cyber security, additional components can add risk by increasing a system's attack surface if they are not verified and securely implemented.

Risks

Key question: How do you ensure that associated software and dependencies are not introducing additional vulnerabilities?

AI systems often rely on numerous libraries and tools, which increases complexity and makes it harder to ensure all components are secure. Even widely used tools and components can pose serious security risks.

Conventional threats (such as typosquatting and name-confusion attacks) remain relevant and should be considered when assessing risk for AI software.

Case study: Vulnerabilities in AI libraries

Several major AI frameworks and libraries have been found to have critical vulnerabilities that could enable deserialization or similar attacks. When exploited, these flaws can lead to token theft or even full remote code execution. In 2024, supply chain compromise of a published AI library led to users unknowingly installing cryptocurrency mining malware when installing the package.

Mitigations

Organisations should continue to validate all components of AI and ML systems as they would with any other piece of software. Steps include malware scanning, integrity validation through checksums or signatures, dynamic and static application testing, and least-privilege practices.

Audit software components and maintain an SBOM to ensure known vulnerabilities are mitigated at deployment and throughout the system's life cycle.

AI infrastructure and hardware

AI systems have unique infrastructure and hardware requirements that can affect their security.

Beyond high computational demands, AI systems more heavily rely on a GPU (graphics processing unit) or even AI-specific accelerator devices.

Risks

Key question: How do you minimise the attack surface introduced by AI infrastructure and hardware?

Most risks to ML infrastructure and hardware mirror those found in conventional cyber systems, as they use broadly similar hardware. As with similar components, AI-specific accelerator devices introduce additional attack surfaces through their drivers, firmware and related components.

As machine-to-machine systems become more common, new communication protocols are being introduced. These may add complexity to systems, such as routing and firewall management.

Mitigations

Treat infrastructure and hardware for AI systems like any other integrated technology. Follow existing security processes and practices to ensure hardware is free from malicious content and appropriately segmented within the network. This includes enforcing signed drivers and firmware, enabling verified boot, logging, and using separate management networks with their own authentication and auditing.

Third-party services

Due to the resource demands and complexity of AI systems, third-party services are often used to support some or all components and services. This can range from using cloud-based data hosting, to developing and managing the entire AI system.

While third-party providers offer many advantages, they also come with additional risks and responsibilities.

Risks

Key question: How do you avoid risk from a third party’s poor cyber hygiene?

Third-party providers of AI and ML tools, platforms and services can introduce vulnerabilities into an organisation's supply chain. These vulnerabilities may be from their own supply chain, a cyber incident, or shared resources with other tenants.

Malicious actors can potentially use these vectors to compromise customers’ systems if they are not well managed. Working with third parties can also create ambiguity around shared responsibilities that could lead to gaps in cyber security.

Mitigations

Perform thorough assessments of third-party vendors, including reviewing their security practices, vulnerability management processes, and overall track record. Ongoing monitoring should ensure adherence to strong security practices.

Vendors should demonstrate a commitment to security and transparency across their products, services, systems and supply chain. Consider cyber security requirements for third parties early and include them in contracts. This could include restricting use of your data for training, defining separate cloud residencies, granting audit rights, setting continuity requirements, and establishing shared responsibility models.

Conclusion

AI and ML can introduce distinct cyber security challenges to a supply chain, particularly due to their reliance on a complex ecosystem of models, data, libraries and cloud infrastructure. Without proper safeguards, threats like poisoned data, backdoors and malicious code can compromise these systems and lead to further exploitation.

To mitigate these risks, organisations should adopt secure supply chain practices that consider the unique risks that AI and ML systems introduce. By combining cyber security best practices and the mitigations in this guidance as part of a defence-in-depth strategy, organisations can reduce their attack surface and strengthen the overall resilience of their AI and ML supply chain.

More information

- Artificial intelligence
- Managing cyber supply chains
- Choosing Secure and Verifiable Technologies
- MITRE ATLAS Matrix
- NIST AI Risk Management Framework
- US CSCC Supplier, Products, and Services Threat Evaluation Report (to include Artificial Intelligence Risks and Mitigations)

Disclaimer

The material in this guide is of a general nature and should not be regarded as legal advice or relied on for assistance in any particular circumstance or emergency situation. In any important matter, you should seek appropriate independent professional advice in relation to your own circumstances.

The Commonwealth accepts no responsibility or liability for any damage, loss or expense incurred as a result of the reliance on information contained in this guide.

Copyright

Use of the Coat of Arms

Australian Government
Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE

ACSC Australian Cyber Security Centre