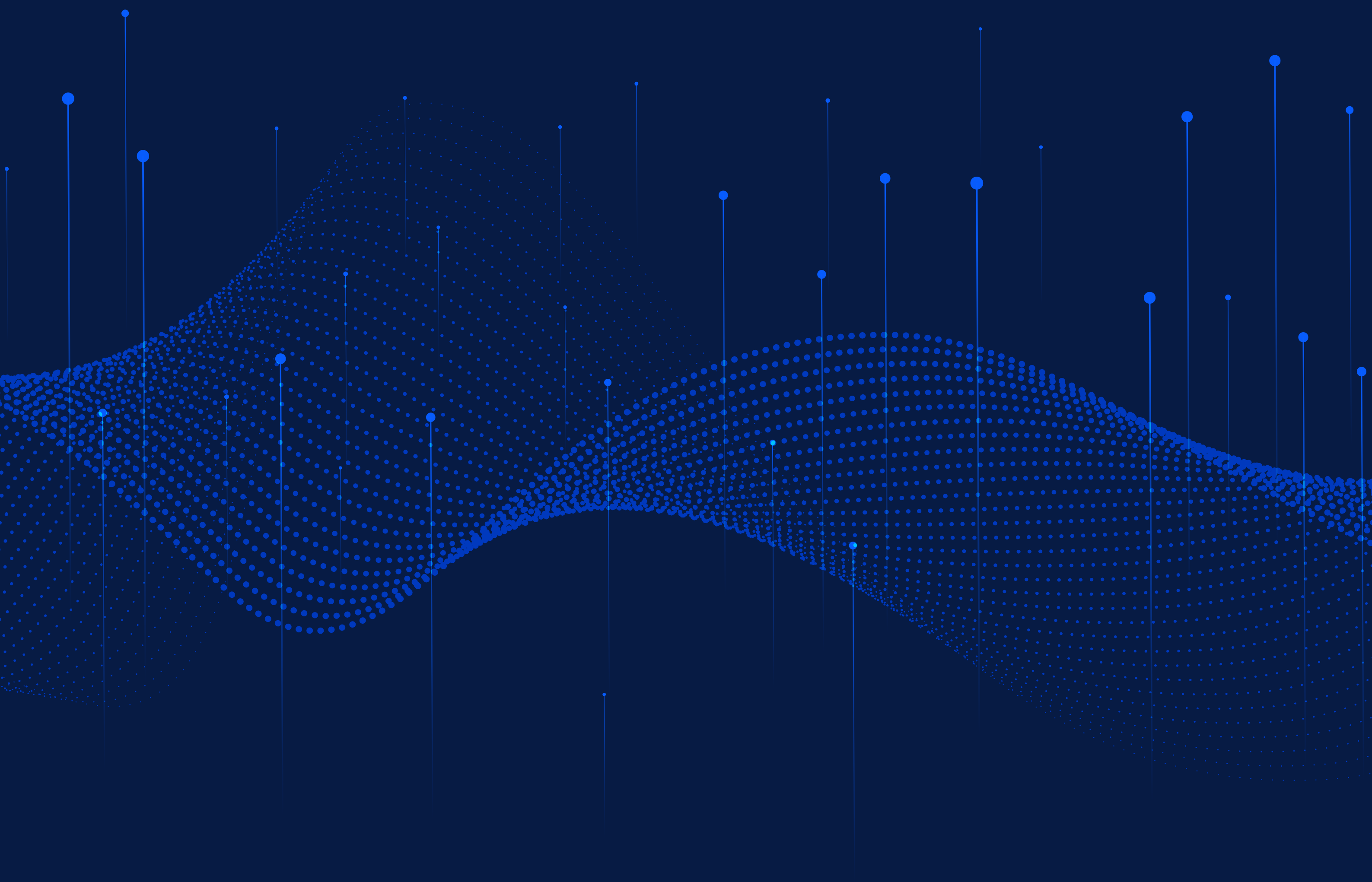


Engaging with Artificial Intelligence (AI)

Australian Government
Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC Australian Cyber Security Centre

CYBERSECURITY & INFRASTRUCTURE SECURITY AGENCY
CISA

DEPARTMENT OF JUSTICE
FEDERAL BUREAU OF INVESTIGATION
FIDELITY BRAVERY INTEGRITY

NATIONAL SECURITY AGENCY
UNITED STATES OF AMERICA

National Cyber Security Centre
a part of GCHQ

Communications Security Establishment
Canadian Centre for Cyber Security

Centre de la sécurité des télécommunications
Centre canadien pour la cybersécurité

National Cyber Security Centre
PART OF THE GCSB

certnz

Federal Office for Information Security

INCD
Israel National Cyber Directorate

NISC
内閣サイバーセキュリティセンター
National center of Incident readiness and Strategy for Cybersecurity

NSM
NORWEGIAN NATIONAL CYBER SECURITY CENTRE

CSA
SINGAPORE

NATIONAL CYBER SECURITY CENTRE SWEDEN

Contents

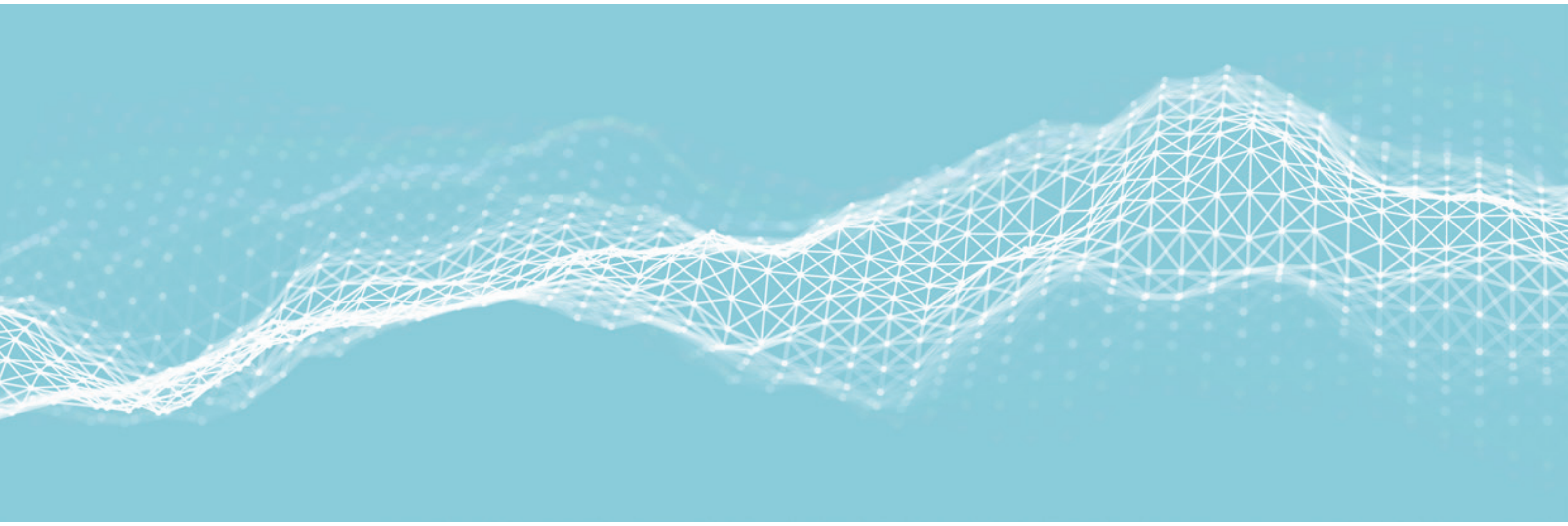
Introduction

The purpose of this publication is to provide organisations with guidance on how to use AI systems securely. The paper summarises some important threats related to AI systems and prompts organisations to consider steps they can take to engage with AI while managing risk. It provides mitigations to assist both organisations that use self-hosted and third-party hosted AI systems.

This publication was developed by the Australian Signals Directorate's Australian Cyber Security Centre (ASD's ACSC) in collaboration with the following international partners:

- United States (US) Cybersecurity and Infrastructure Security Agency (CISA), the Federal Bureau of Investigation (FBI) and the National Security Agency (NSA)
- United Kingdom (UK) National Cyber Security Centre (NCSC-UK)
- Canadian Centre for Cyber Security (CCCS)
- New Zealand National Cyber Security Centre (NCSC-NZ) and CERT NZ
- Germany Federal Office for Information Security (BSI)
- Israel National Cyber Directorate (INCD)
- Japan National Center of Incident Readiness and Strategy for Cybersecurity (NISC) and the Secretariat of Science, Technology and Innovation Policy, Cabinet Office
- Norway National Cyber Security Centre (NCSC-NO)
- Singapore Cyber Security Agency (CSA)
- Sweden National Cybersecurity Center

The guidance within this publication is focused on using AI systems securely rather than developing secure AI systems. The authoring agencies encourage developers of AI systems to refer to the joint Guidelines for Secure AI System Development.



What is AI?

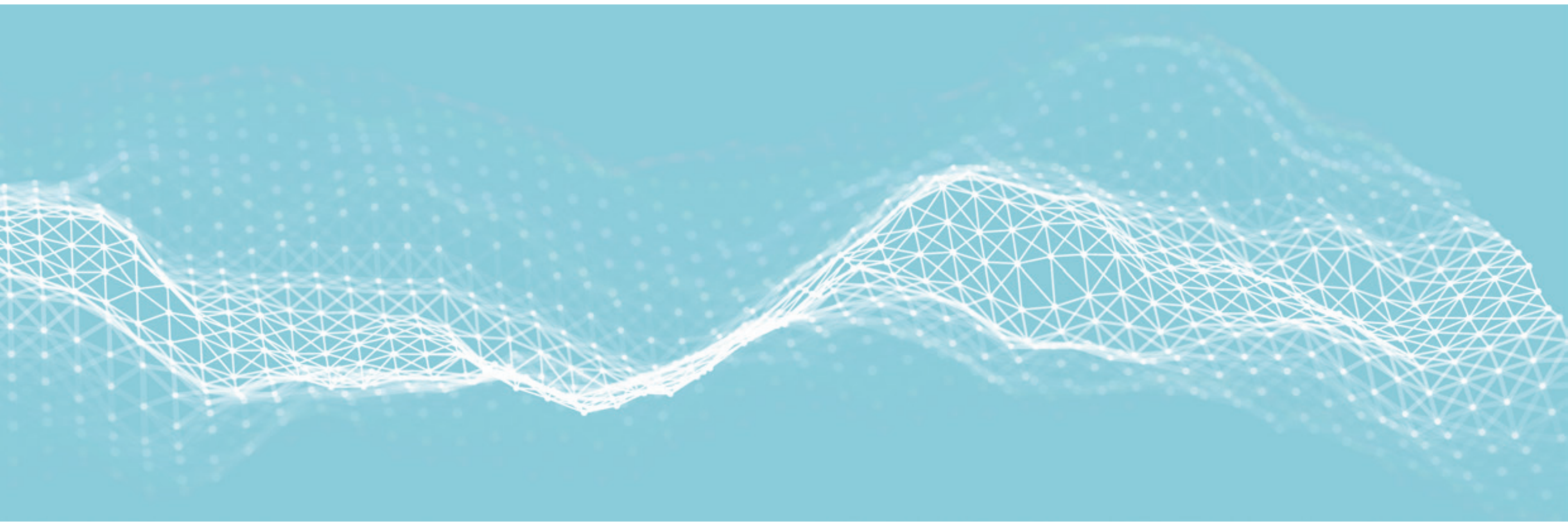
AI is the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making and translation between languages. Modern AI is usually built using machine learning algorithms.

AI has several sub-fields of importance that include but are not limited to:

- **Machine learning** describes software components (models) that allow computers to recognise and bring context to patterns in data without the rules having to be explicitly programmed by a human. Machine learning applications can generate predictions, recommendations, or decisions based on statistical reasoning.
- **Natural language processing** analyses and derives information from human language sources, including text, image, video and audio data. Natural language processing applications are commonly used for language classification and interpretation. Many natural language processing applications not only process natural language but also generate content that mimics it.
- **Generative AI** refers to systems that use data models to generate new examples of content such as text, images, audio, code and other data modalities. Generative AI applications are typically trained on large amounts of real-world data and can approximate human generated content from prompts, even prompts that are limited or non-specific.

AI systems are among the fastest growing applications globally. AI drives internet searching, satellite navigation, and recommendation systems. AI is also increasingly used to handle activities traditionally undertaken by humans such as sorting large data sets, automating routine tasks, creative endeavours and augmenting business activities such as customer engagement, logistics, medical diagnosis, and cyber security.

Organisations from all sectors are exploring opportunities to improve their operations with AI. While AI has the potential to increase efficiency and lower costs, it can also intentionally or inadvertently cause harm. For this reason, government, academia and industry have a role to play in managing the risks associated with this technology, including through research, regulation, policy and governance.



Challenges when engaging with AI

Like all digital systems, AI presents both opportunities and threats. To take advantage of the benefits of AI securely, all stakeholders involved with these systems (e.g. programmers, end users, senior executives, analysts, marketers) should take some time to understand what threats apply to them and how those threats can be mitigated.

Some common AI related threats are outlined below. These threats are not presented to dissuade AI use, but rather to assist all AI stakeholders to engage with AI securely. Cyber security mitigations that can be used to secure against these AI threats are covered in a later section of this publication. For more information on AI-specific threats, adversary tactics and how they can be risk-managed visit MITRE ATLAS and the US National Institute of Standards and Technology's (NIST) AI Risk Management Framework.

1. Data Poisoning of an AI Model

NIST outlines that the data-driven approach of machine learning introduces security and privacy challenges that are different from other systems (see NIST's Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations). These challenges include the potential for manipulation of training data and exploitation of model vulnerabilities (also known as adversarial AI), which can adversely affect the performance of machine learning tools and systems.

One method of adversarial manipulation is data poisoning. Data poisoning involves manipulating an AI model's training data so that the model learns incorrect patterns and may misclassify data or produce inaccurate, biased or malicious outputs. Any organisational function that relies on the integrity of the AI system's outputs could be negatively impacted by data poisoning. An AI model's training data could be manipulated by inserting new data or modifying existing data; or the training data could be taken from a source that was poisoned to begin with. Data poisoning may also occur in the model's fine-tuning process. An AI model that receives feedback to determine if it has correctly performed its function could be manipulated by poor quality or incorrect feedback.

Case Study: Tay Chatbot Poisoning

In 2016, Microsoft trialled a Twitter chatbot called "Tay" that leveraged machine learning. The chatbot used its conversations with users to train itself and adapt its interactions, leaving it vulnerable to a data poisoning attack. Users tweeted abusive language at Tay with the intention of inserting abusive material into Tay's training data therefore poisoning the AI model when it was retrained. As a result, Tay began using abusive language towards other users.

2. Input manipulation attacks – Prompt injection and adversarial examples

Prompt injection is an input manipulation attack that attempts to insert malicious instructions or hidden commands into an AI system. Prompt injection can allow a malicious actor to hijack the AI model's output and jailbreak the AI system. In doing so, the malicious actor can evade content filters and other safeguards restricting the AI system's functionality.

Case Study: DAN prompt injection

A widely reported example of prompt injection leading to jailbreaking an AI system is the "Do Anything Now" (DAN) prompt. Users of ChatGPT have discovered various ways to prompt ChatGPT to assume an identity named "DAN" that is not subject to the system's usual safety restrictions. OpenAI's efforts to address this case of prompt injection have been bypassed on several occasions by new iterations of the DAN prompt, highlighting the challenges of enforcing safety restrictions on AI systems.

Another type of input manipulation attack is known as 'adversarial examples'. In the context of AI, adversarial examples are the crafting of specialised inputs that, when given to an AI, intentionally cause it to produce incorrect outputs, such as misclassifications. Inputs can be crafted to pass a confidence test, return an incorrect result or bypass a detection mechanism. Note that, in adversarial examples, inputs to the AI are manipulated while it is in use, rather than when it is being trained. For example, consider a music sharing service that requires user submitted music to pass an AI powered copyright check before it is published. In an adversarial example attack, a user might slightly speed up a copyright protected song so that it passes the AI powered copyright check while remaining recognisable to listeners.

3. Generative AI hallucinations

Outputs generated by an AI system may not always be accurate or factually correct. Generative AI systems are known to hallucinate information that is not factually correct. Organisational functions that rely on the accuracy of generative AI outputs could be negatively impacted by hallucinations, unless appropriate mitigations are implemented.

Case study: Southern District of New York Legal Filing

In 2023, a district judge in the Southern District of New York reportedly found that a legal brief submitted to him contained at least six hallucinated cases. The lawyer who submitted the brief attributed the hallucinated cases to research he undertook with ChatGPT, admitting in an affidavit that he "was unaware of the possibility that its content could be false."

4. Privacy and intellectual property concerns

AI systems may also present a challenge to ensuring the security of sensitive data an organisation holds, including customers' personal data and intellectual property.

Organisations should be cautious with what information they, and their personnel, provide to generative AI systems. Information given to these systems may be incorporated into the system's training data and could inform outputs to prompts from non-organisational users.

For further information on applying privacy principles to generative AI technologies, visit the Office of the Privacy Commissioner of Canada's Principles for responsible, trustworthy and privacy-protective generative AI technologies.

5. Model stealing attack

A model stealing attack involves a malicious actor providing inputs to an AI system and using the outputs to create an approximate replica of it. AI models can require a significant investment to create, and the prospect of model stealing is a serious intellectual property concern. For example, consider an insurance company that has developed an AI model to provide customers with insurance quotes. If a competitor were to query this model to the extent that it could create a replica of it, it could benefit from the investment that went into creating the model, without sharing in its development costs.

Similarly, there are cases where prompts have led generative AI models to divulge training data. The exfiltration of training data can be a serious privacy concern for models that are trained on sensitive data, including data that contains personally identifiable information. It is also a serious intellectual property concern for organisations that seek to maintain the confidentiality of their training data sets.

Case study: ChatGPT memorised training data extraction

In November 2023, a team of researchers published the outcomes of their attempts to extract memorised training data from AI language models. One of the applications the researchers experimented with was ChatGPT. In the case of ChatGPT, the researchers found that prompting the model to repeat a word forever led the model to divulge training data at a rate much higher than when behaving as normal[1]. The extracted training data included personally identifiable information.

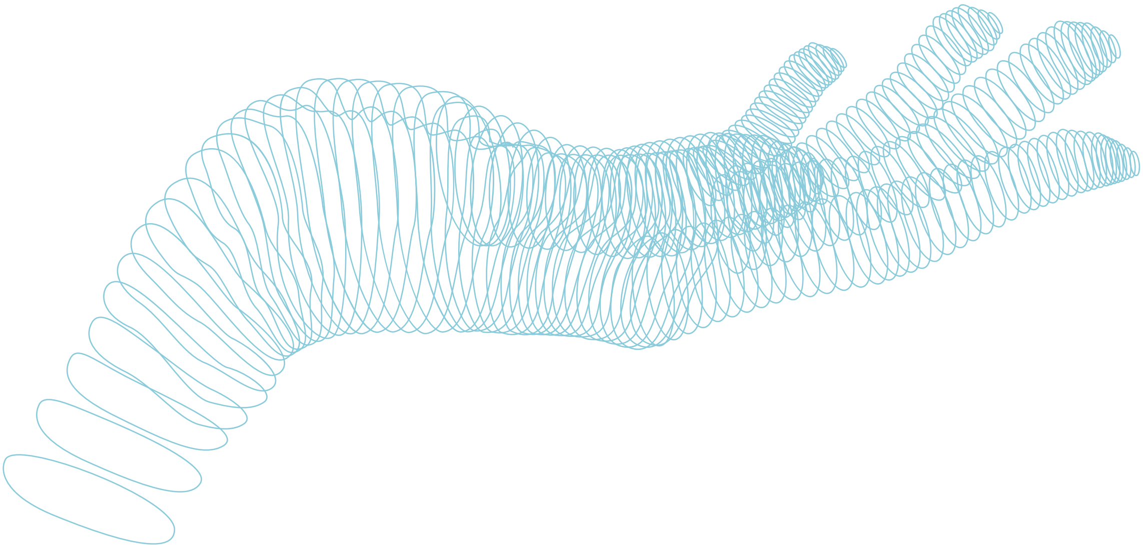
[1] Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A.F., Ippolito, D., Choquette-Choo, C.A., Wallace, E., Tramèr, F. and Lee, K., 2023. Scalable extraction of training data from (production) language models. arXiv preprint arXiv:2311.17035.

Mitigation considerations for organisations

AI technologies are distinctive in their speed of innovation and scope of impact. As a result, it is important that organisations that use, or are considering using, AI systems, consider their cyber security implications. This includes evaluating the AI system's benefits and risks within the organisation's context. The questions below are intended to prompt organisations to consider how they can use AI systems securely. They include a number of cyber security mitigations that may be relevant to their use of both their self-hosted AI systems and third-party AI systems. As an emerging technology, there are limited regulations to ensure AI systems are secure. In the absence of a robust regulatory framework, it is important organisations carefully consider the risks associated with any AI system they are considering using. As AI is evolving quickly, the below mitigation considerations may need to be revisited on an ongoing basis.

Has your organisation implemented the cyber security frameworks relevant to its jurisdiction?

Your organisation's AI systems would benefit from many of the same cyber security mitigations that you have implemented to protect your organisation's other systems. Start by ensuring that you have implemented the cyber security mitigations recommended in the framework that is relevant to your jurisdiction. The "Further Reading" section below includes links to several cyber security frameworks developed by the authors of this publication.



How will the system affect your organisation's privacy and data protection obligations?

Consider how the AI system collects, processes and stores data, and how this may impact your organisation's privacy and data protection obligations.

- AI systems are often hosted in the cloud and may send data between different regions. Ensure that any AI system your organisation uses can meet your data residency or sovereignty obligations.
- If using a third-party AI system, understand if your organisation's inputs will be used to retrain the AI system's model. Consider using private versions of the system, if available.
- If using a third-party AI system, ensure your organisation is aware of how your data will be managed in the event your commercial agreement with the third-party ends. This information is typically outlined in the vendor's privacy policy or terms of service.
- If your AI system handles private data, consider if there are privacy enhancing technologies you can employ to protect that data.

Does your organisation enforce multi-factor authentication?

Require phishing-resistant multi-factor authentication, for example, FIDO2 security keys, to access your organisation's AI systems, including any repositories that hold training data. Multi-factor authentication protects against unauthorised access to your organisation's systems and resources. Unauthorised access could facilitate several attacks including data poisoning and model theft.

How will your organisation manage privileged access to the AI system?

Grant privileges based on the need-to-know principle and the principle of least privilege. For example, limit the number of accounts with access to the AI's development and production environments and limit access to repositories that hold the AI model's training data. Require that privileged accounts are routinely revalidated and disabled after a set period of inactivity. Restricting privileged access to the system mitigates several threats, including data poisoning and model theft.

How will your organisation manage backups of the AI system?

Maintain backups of your AI model and training data. Backups will assist your organisation to recover if your AI system is affected by an incident. For example, if your organisation is affected by a data poisoning attack, maintaining backups would allow it to restore an unaffected copy of its training data and retrain its model. It should be noted that backing up your AI model and its training data can be resource intensive.

Can your organisation implement a trial of the AI system?

Trialling an AI system can be an effective way to understand how the system can integrate with your organisation's cyber security systems and tools, for example, firewalls, gateways, extended detection and response tools, logging and monitoring systems. A trial can also help your organisation to test the limits and constraints of the system in a low-stakes environment. Before implementing a trial, consider its scope and success criteria.

Is the AI system secure-by-design, including its supply chain?

Seek to use vendors that are transparent about how they have developed and tested their AI systems. Consider if the AI system has applied the guidelines recommended in the NCSC-UK's Guidelines for secure AI system development.

The supply chains of AI systems can be complex and, as a result, are likely to carry inherent risks. Conducting a supply chain evaluation can help you to identify and manage these risks.

If your organisation is involved in training the AI system it uses, consider the supply chain of foundational training data and fine-tuning data as well, to aid in preventing data poisoning. The security of the data and model parameters is critical.

Does your organisation understand the limits and constraints of the AI system?

AI systems can be incredibly complex. While it is often not practical, or possible, to understand the intricacies of how AI systems work, it is still helpful to understand their general limits and constraints. For example, is the AI system prone to hallucinations? If the system is involved in data classification, what is its rate of false positives and false negatives? Understanding the system's limits and constraints will assist your organisation to account for them in its processes.

Does your organisation have suitably qualified staff to ensure the AI system is set-up, maintained and used securely?

Ensure that your organisation is adequately resourced to securely set-up, maintain and use the AI system.

Consider which staff would be interacting with the AI system, what these staff would be required to know to interact with the system securely and how this knowledge can be developed.

Staff that use the system should be trained on what data can and cannot be input to the system, for example, personally identifiable information or the organisation's intellectual property. Staff should also be trained on the extent to which the system's outputs can be relied upon and any organisational processes for output validation.

Does your organisation conduct health checks of your AI system?

Conduct periodic health checks of AI systems to detect data drift and ensure the system is working efficiently and as intended. Data drift describes when the data an AI system encounters in the real world differs from the data that system was trained on. Data drift can lead to a degradation in the AI system's performance. It typically occurs over time, as the environment the AI system operates in changes. Periodically updating the system's training data with data received during normal usage can mitigate data drift. For more information on secure operation and maintenance of your organisation's AI system visit the joint publication Guidelines for Secure AI System Development.

Does your organisation enforce logging and monitoring?

Consider how your organisation would detect anomalies or malicious activity associated with the AI system.

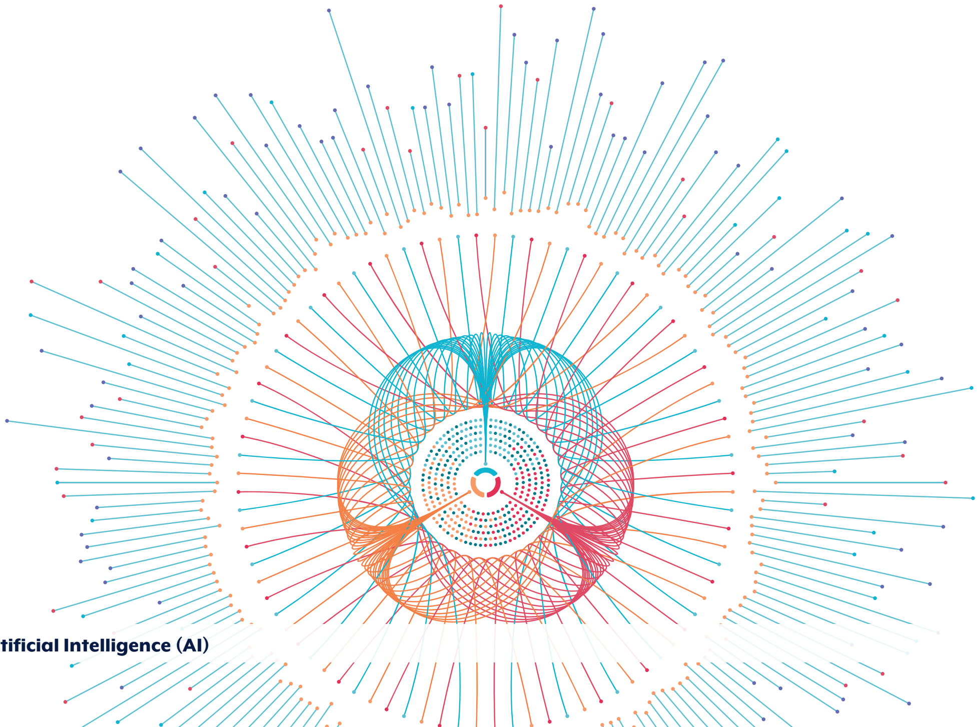
- Log and monitor outputs from the AI system to detect any change in behaviour or performance that may indicate a compromise or data drift.
- Log and monitor inputs to the AI system to ensure compliance obligations are met and to aid investigation and remediation efforts in the event of an incident.
- Log and monitor the network and endpoints that host your AI system to detect attempts to access, modify or copy system data.
- Log and monitor logins to repositories that hold training data, the AI system's development and production environments and backups. Consider how any logging and monitoring tools your organisation employs may integrate with your AI system.
- Log and monitor for high frequency, repetitive prompts. These can be a sign of automated prompt injection attacks.
- Establish a baseline of the AI system's activity to assist your organisation to determine when logged events are anomalous.

What will your organisation do if something goes wrong with the AI system?

Consider how your organisation may be impacted if an incident or error affects the AI system so that you can implement proportionate mitigations and contingencies.

If you are using a third-party AI system, familiarise yourself with any up-time or availability commitments the vendor has made. Ensure that vendor and customer responsibilities regarding incident management are clearly defined in the service contract.

Ensure that your organisation's incident response plan accounts for issues arising from, or to, its AI systems and consider how business continuity can be achieved in the event of a serious incident. Your incident response plan should also clearly define the roles and responsibilities that are critical to addressing any incident that affects the AI system.



Further Reading

ASD's Cyber Supply Chain Risk Management

Guidance published by ASD to help organisations manage risk in their cyber supply chain.

ASD's Essential Eight

ASD has developed prioritised mitigation strategies, in the form of the Strategies to Mitigate Cyber Security Incidents, to help organisations protect themselves against various cyber threats. The most effective of these mitigation strategies are the Essential Eight. The Essential Eight has been designed to protect organisations' internet-connected information technology networks.

ASD's Ethical AI framework

ASD has developed a framework that incorporates a set of ethical principles which govern how AI is used at ASD.

ASD's Information Security Manual

ASD produces the Information Security Manual. The purpose of the Information Security Manual is to outline a cyber security framework that an organisation can apply, using their risk management framework, to protect their systems and data from cyber threats. The Information Security Manual is intended for Chief Information Security Officers, Chief Information Officers, cyber security professionals and information technology managers.

BSI's AI Cloud Service Compliance Criteria Catalogue (AIC4)

BSI's AI Cloud Compliance Criteria Catalogue provides AI-specific criteria, which enable evaluation of the security of an AI service across its lifecycle.

BSI's Large Language Models: Opportunities and Risks for Industry and Authorities

Document produced by BSI for companies, authorities and developers who want to learn more about the opportunities and risks of developing, deploying and/or using LLMs.

CCCS Principles for responsible, trustworthy and privacy-protective generative AI technologies

The Office of the Privacy Commissioner of Canada has published guidance to help organisations developing, providing or using generative AI to apply key Canadian privacy principles.

CERT NZ's Top online security tips for your business

Guidance that includes cyber security mitigation strategies as well as why they matter and how to implement them.

Hiroshima AI Process Comprehensive Policy Framework

This is the first international framework that includes guiding principles and a code of conduct aimed at promoting the safe, secure and trustworthy development of advanced AI systems. The policy framework was successfully agreed upon at the G7 Digital & Tech Ministers' Meeting in December 2023 and was endorsed by the G7 Leaders in the same month.

The Hiroshima AI Process was launched in May 2023, following the Leaders' direction at the G7 Hiroshima Summit in Japan.

MITRE ATLAS

MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems) is a globally accessible, living knowledge base of adversary tactics and techniques based on real-world attack observations and realistic demonstrations from AI red teams and security groups.

NIST Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations

This NIST report on AI develops a taxonomy of attacks and mitigations and defines terminology in the field of adversarial machine learning. Taken together, the taxonomy and terminology are meant to inform other standards and future practice guides for assessing and managing the security of AI systems by establishing a common language for understanding the rapidly developing adversarial machine learning landscape.

NIST AI Risk Management Framework

In collaboration with the private and public sectors, NIST has developed a framework to better manage risks to individuals, organizations, and society associated with AI. It is intended for voluntary use and to improve the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems.

NIST Cybersecurity Framework

The NIST Cybersecurity Framework consists of standards, guidelines and best practices to manage cyber security risk. It is comprised of three parts: the Framework Core, the Framework Implementation Tiers, and the Framework Profiles. Each framework component reinforces the connection between business/mission drivers and cyber security activities.

NCSC NZ Cyber Security Framework

NCSC NZ's cyber security framework sets out how NCSC NZ thinks, talks about, and organises its cyber security efforts. Its five functions and twenty-five security objectives represent the breadth of work needed to secure an organisation in New Zealand.

NCSC-NZ Interim Generative AI guidance for the public service

The New Zealand Government has published interim guidance on the use of generative AI in the public service. The guidance includes 10 "do's" for trustworthy use of generative AI in the public service.

NCSC-UK 10 Steps to Cyber Security

The NCSC-UK's 10 steps to cyber security provides a summary of the NCSC-UK's advice for medium to large organisations. It aims to help organisations manage their cyber security risks by breaking down the task of protecting the organisation into 10 components.

NCSC-UK Guidelines for Secure AI System Development

Guidelines co-sealed by Australia, Canada, New Zealand, the United Kingdom, the United States and a number of international partners for providers of any systems that use artificial intelligence (AI), whether those systems have been created from scratch or built on top of tools and services provided by others.

NCSC-UK Principles for the security of machine learning

These principles aim to be wide reaching and applicable to anyone developing, deploying or operating a system with a machine learning (ML) component. They are not a comprehensive assurance framework to grade a system or workflow, and do not provide a checklist. Instead, they provide context and structure to help scientists, engineers, decision makers and risk owners make educated decisions about system design and development processes, helping to assess the specific threats to a system.

OWASP Machine Learning Security Top 10

The OWASP Machine Learning Security Top 10 project delivers an overview of the top 10 security issues relating to machine learning systems.

US Department of Defense 2023 Data, Analytics, and Artificial Intelligence Adoption Strategy

This strategy's approach embraces the need for speed, agility, learning, and responsibility. Pursuing this agile approach and focusing activities on the goals outlined in this strategy will allow the US Department of Defense to adopt data, analytics, and AI-enabled capabilities at the pace and scale required to build enduring decision advantage.

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

For more information, or to report a cyber security incident, contact us:
cyber.gov.au | 1300 CYBER1 (1300 292 371)

Australian Government
Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC Australian Cyber Security Centre