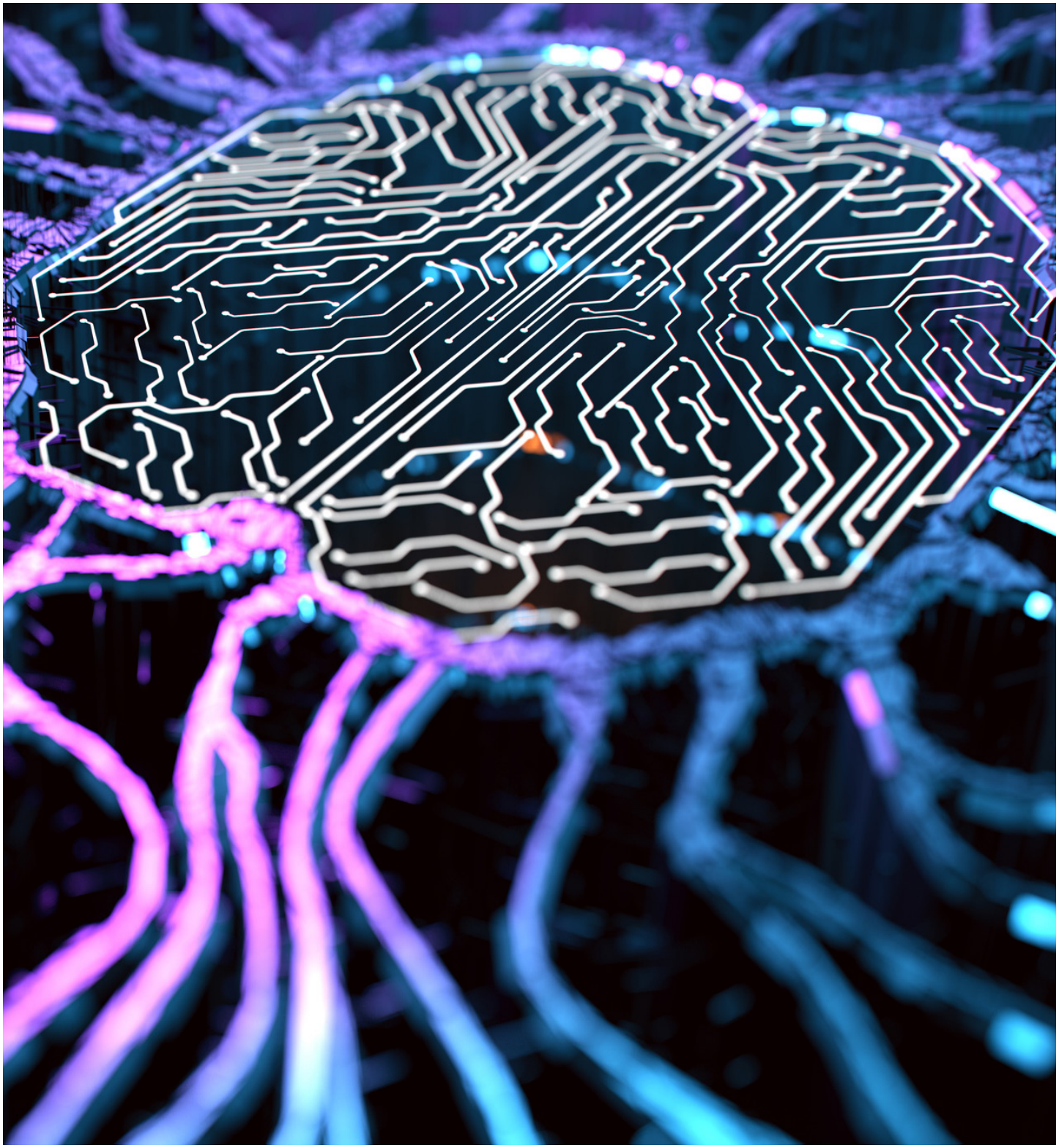
# Guidelines for secure AI system development



National Cyber Security Centre
a part of GCHQ

CISA
Cybersecurity & Infrastructure Security Agency

Australian Government
Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC Australian Cyber Security Centre

Communications Security Establishment
Canadian Centre for Cyber Security

Centre de la sécurité des télécommunications
Centre canadien pour la cybersécurité

CSIRT
Equipo de Respuesta ante Incidentes de Seguridad Informática

National Cyber and Information Security Agency

Republic of Estonia
Information System Authority

Department of Justice
Federal Bureau of Investigation

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Federal Office for Information Security

INCD
Israel National Cyber Directorate

Agenzia per la Cybersicurezza Nazionale
ACN
Sapientia Adiuvat

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

National Cyber Security Centre

NITDA

NSM
NORWEGIAN NATIONAL CYBER SECURITY CENTRE

National Security Agency
United States of America

NASK

국가정보원
National Intelligence Service

Ministerstwo Cyfryzacji

CSA SINGAPORE
Cyber Security Agency of Singapore

# About this document

This document is published by the UK National Cyber Security Centre (NCSC), the US Cybersecurity and Infrastructure Security Agency (CISA), and the following international partners:

- National Security Agency (NSA)
- Federal Bureau of Investigation (FBI)
- Australian Signals Directorate's Australian Cyber Security Centre (ACSC)
- Canadian Centre for Cyber Security (CCCS)
- New Zealand National Cyber Security Centre (NCSC-NZ)
- Chile's Government CSIRT
- National Cyber and Information Security Agency of the Czech Republic (NUKIB)
- Information System Authority of Estonia (RIA)
- National Cyber Security Centre of Estonia (NCSC-EE)
- French Cybersecurity Agency (ANSSI)
- Germany's Federal Office for Information Security (BSI)
- Israeli National Cyber Directorate (INCD)
- Italian National Cybersecurity Agency (ACN)
- Japan's National center of Incident readiness and Strategy for Cybersecurity (NISC)
- Japan's Secretariat of Science, Technology and Innovation Policy, Cabinet Office
- Nigeria's National Information Technology Development Agency (NITDA)
- Norwegian National Cyber Security Centre (NCSC-NO)
- Poland Ministry of Digital Affairs
- Poland's NASK National Research Institute (NASK)
- Republic of Korea National Intelligence Service (NIS)
- Cyber Security Agency of Singapore (CSA)

# Acknowledgements

The following organisations contributed to the development of these guidelines:

- Alan Turing Institute
- Amazon
- Anthropic
- Databricks
- Georgetown University's Center for Security and Emerging Technology
- Google
- Google DeepMind
- Hugging Face
- IBM
- Imbue
- Inflection
- Microsoft
- OpenAI
- Palantir
- RAND
- Scale AI
- Software Engineering Institute at Carnegie Mellon University
- Stanford Center for AI Safety
- Stanford Program on Geopolitics, Technology and Governance

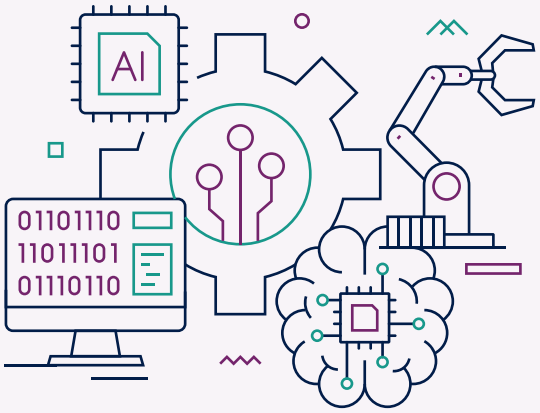
# Disclaimer

The information in this document is provided "as is" by the NCSC and the authoring organisations who shall not be liable for any loss, injury or damage of any kind caused by its use save as may be required by law. The information in this document does not constitute or imply endorsement or recommendation of any third party organisation, product, or service by the NCSC and authoring agencies. Links and references to websites and third party materials are provided for information only and do not represent endorsement or recommendation of such resources over others.

This document is made available on a TLP:CLEAR basis (https://www.first.org/tlp/).

# Contents





# Executive summary

**This document recommends guidelines 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. Implementing these guidelines will help providers build AI systems that function as intended, are available when needed, and work without revealing sensitive data to unauthorised parties.**

This document is aimed primarily at providers of AI systems who are using models hosted by an organisation, or are using external application programming interfaces (APIs). We urge **all** stakeholders (including data scientists, developers, managers, decision-makers and risk owners) to read these guidelines to help them make informed decisions about the **design**, **development**, **deployment** and **operation** of their AI systems.

## About the guidelines

AI systems have the potential to bring many benefits to society. However, for the opportunities of AI to be fully realised, it must be developed, deployed and operated in a secure and responsible way.

AI systems are subject to novel security vulnerabilities that need to be considered alongside standard cyber security threats. When the pace of development is high – as is the case with AI – security can often be a secondary consideration. Security must be a core requirement, not just in the development phase, but throughout the life cycle of the system.

For this reason, the guidelines are broken down into four key areas within the AI system development life cycle: **secure design**, **secure development**, **secure deployment**, and **secure operation and maintenance**. For each section we suggest considerations and mitigations that will help reduce the overall risk to an organisational AI system development process.

1. **Secure design**
   This section contains guidelines that apply to the design stage of the AI system development life cycle. It covers understanding risks and threat modelling, as well as specific topics and trade-offs to consider on system and model design.

2. **Secure development**
   This section contains guidelines that apply to the development stage of the AI system development life cycle, including supply chain security, documentation, and asset and technical debt management.

3. **Secure deployment**
   This section contains guidelines that apply to the deployment stage of the AI system development life cycle, including protecting infrastructure and models from compromise, threat or loss, developing incident management processes, and responsible release.

4. **Secure operation and maintenance**
   This section contains guidelines that apply to the secure operation and maintenance stage of the AI system development life cycle. It provides guidelines on actions particularly relevant once a system has been deployed, including logging and monitoring, update management and information sharing.

The guidelines follow a 'secure by default' approach, and are aligned closely to practices defined in the NCSC's Secure development and deployment guidance, NIST's Secure Software Development Framework, and 'secure by design principles' published by CISA, the NCSC and international cyber agencies. They prioritise:

- taking ownership of security outcomes for customers
- embracing radical transparency and accountability
- building organisational structure and leadership so secure by design is a top business priority

# Introduction

Artificial intelligence (AI) systems have the potential to bring many benefits to society. However, for the opportunities of AI to be fully realised, it must be developed, deployed and operated in a secure and responsible way. Cyber security is a necessary precondition for the safety, resilience, privacy, fairness, efficacy and reliability of AI systems.

However, AI systems are subject to novel security vulnerabilities that need to be considered alongside standard cyber security threats. When the pace of development is high – as is the case with AI – security can often be a secondary consideration. Security must be a core requirement, not just in the development phase, but throughout the life cycle of the system.

**This document recommends guidelines for providers[1] of any systems that use AI, whether those systems have been created from scratch or built on top of tools and services provided by others. Implementing these guidelines will help providers build AI systems that function as intended, are available when needed, and work without revealing sensitive data to unauthorised parties.**

These guidelines should be considered in conjunction with established cyber security, risk management, and incident response best practice. In particular, we urge providers to follow the 'secure by design'[2] principles developed by the US Cybersecurity and Infrastructure Security Agency (CISA), the UK National Cyber Security Centre (NCSC), and all our international partners. The principles prioritise:

- taking ownership of security outcomes for customers
- embracing radical transparency and accountability
- building organisational structure and leadership so secure by design is a top business priority.

Following 'secure by design' principles requires significant resources throughout a system's life cycle. It means developers must invest in prioritising **features**, **mechanisms**, and **implementation** of tools that protect customers at each layer of the system design, and across all stages of the development life cycle. Doing this will prevent costly redesigns later, as well as safeguarding customers and their data in the near term.

## Why is AI security different?

In this document we use 'AI' to refer specifically to machine learning (ML) applications[3]. All types of ML are in scope. We define ML applications as applications that:

- involve 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
- generate predictions, recommendations, or decisions based on statistical reasoning

As well as existing cyber security threats, AI systems are subject to new types of vulnerabilities. The term 'adversarial machine learning' (AML), is used to describe the exploitation of fundamental vulnerabilities in ML components, including hardware, software, workflows and supply chains. AML enables attackers to cause unintended behaviours in ML systems which can include:

- affecting the model's classification or regression performance
- allowing users to perform unauthorised actions
- extracting sensitive model information

There are many ways to achieve these effects, such as prompt injection attacks in the large language model (LLM) domain, or deliberately corrupting the training data or user feedback (known as 'data poisoning').

# Who should read this document?

This document is aimed primarily at providers of AI systems, whether based on models hosted by an organisation or making use of external application programming interfaces (APIs). However, we urge **all** stakeholders (including data scientists, developers, managers, decision-makers and risk owners) to read these guidelines to help them make informed decisions about the **design**, **deployment** and **operation** of their machine learning AI systems.

That said, not all of the guidelines will be directly applicable to all organisations. The level of sophistication and the methods of attack will vary depending on the adversary targeting the AI system, so the guidelines should be considered alongside your organisation's use cases and threat profile.

# Who is responsible for developing secure AI?

There are often many actors in modern AI supply chains. A simple approach assumes two entities:

- the 'provider' who is responsible for data curation, algorithmic development, design, deployment and maintenance
- the 'user', who provides inputs and receives outputs

While this provider-user approach is used in many applications, it is becoming increasingly uncommon, as providers may look to incorporate software, data, models and/or remote services provided by third parties into their own systems. These complex supply chains make it harder for end users to understand where responsibility for secure AI lies.

Users (whether 'end users', or providers incorporating an external AI component) do not typically have sufficient visibility and/or expertise to fully understand, evaluate or address risks associated with systems they are using. As such, in line with 'secure by design' principles, **providers of AI components should take responsibility for the security outcomes of users further down the supply chain.**

Providers should implement security controls and mitigations where possible within their models, pipelines and/or systems, and where settings are used, implement the most secure option as default. Where risks cannot be mitigated, the provider should be responsible for:

- informing users further down the supply chain of the risks that they and (if applicable) their own users are accepting
- advising them on how to use the component securely

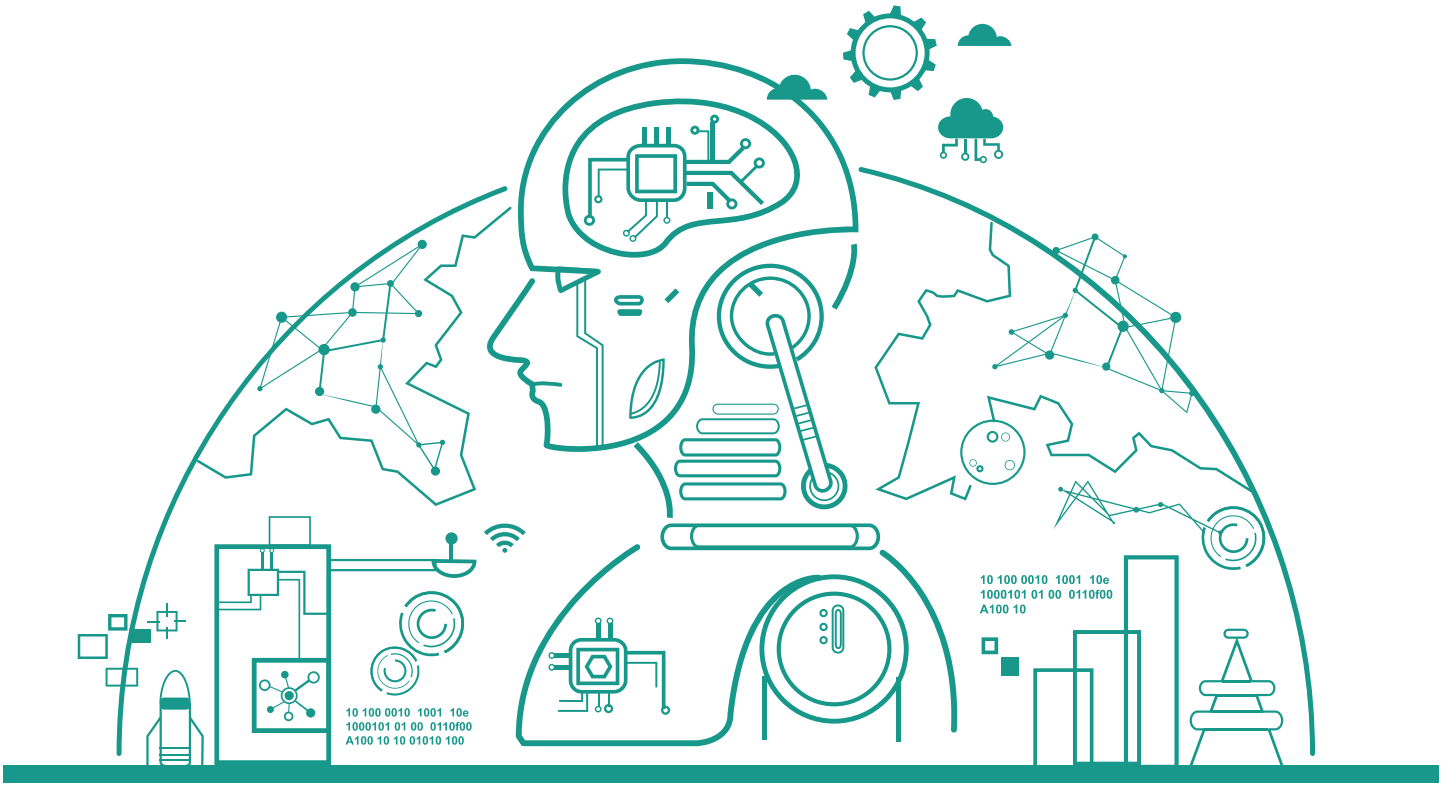
Where system compromise could lead to tangible or widespread physical or reputational damage, significant loss of business operations, leakage of sensitive or confidential information and/or legal implications, AI cyber security risks should be treated as **critical**.

# Guidelines for secure AI system development

The guidelines are broken down into four key areas within the AI system development life cycle: **secure design**, **secure development**, **secure deployment**, and **secure operation and maintenance**. For each area, we suggest considerations and mitigations that will help reduce the overall risk to the organisational AI system development process.

The guidelines set out in this document are aligned closely to software development life cycle practices defined in:

- the NCSC's Secure development and deployment guidance
- the National Institute of Standards and Technology (NIST) Secure Software Development Framework (SSDF)[6]



# 1. Secure design

This section contains guidelines that apply to the **design** stage of the AI system development life cycle. It covers understanding risks and threat modelling, as well as specific topics and trade-offs to consider on system and model design.

### Raise staff awareness of threats and risks 

System owners and senior leaders understand threats to secure AI and their mitigations. Your data scientists and developers maintain an awareness of relevant security threats and failure modes and help risk owners to make informed decisions. You provide users with guidance on the unique security risks facing AI systems (for example, as part of standard InfoSec training) and train developers in secure coding techniques and secure and responsible AI practices.

### Model the threats to your system 

As part of your risk management process, you apply a holistic process to assess the threats to your system, which includes understanding the potential impacts to the system, users, organisations, and wider society if an AI component is compromised or behaves unexpectedly[7]. This process involves assessing the impact of AI-specific threats[8] and documenting your decision making.

You recognise that the sensitivity and types of data used in your system may influence its value as a target to an attacker. Your assessment should consider that some threats may grow as AI systems increasingly become viewed as high value targets, and as AI itself enables new, automated attack vectors.

### Design your system for security as well as functionality and performance 

You are confident that the task at hand is most appropriately addressed using AI. Having determined this, you assess the appropriateness of your AI-specific design choices. You consider your threat model and associated security mitigations alongside functionality, user experience, deployment environment, performance, assurance, oversight, ethical and legal requirements, among other considerations. For example:

- you consider supply chain security when choosing whether to develop in house or use external components, for example:
    - your choice to train a new model, use an existing model (with or without fine-tuning) or access a model via an external API is appropriate to your requirements
    - your choice to work with an external model provider includes a due diligence evaluation of that provider's own security posture
    - if using an external library, you complete a due diligence evaluation (for example, to ensure the library has controls that prevent the system loading untrusted models without immediately exposing themselves to arbitrary code execution[9])
    - you implement scanning and isolation/sandboxing when importing third-party models or serialised weights, which should be treated as untrusted third-party code and could enable remote code execution

  - if using an external API, you apply appropriate controls to data that can be sent to services outside of your organisation's control, such as requiring users to log in and confirm before sending potentially sensitive information
  - you apply appropriate checks and sanitisation of data and inputs; this includes when incorporating user feedback or continuous learning data into your model, recognising that training data defines system behaviour

- you integrate AI software system development into existing secure development and operations best practices; all elements of the AI system are written in appropriate environments using coding practices and languages that reduce or eliminate known classes of vulnerabilities where plausible

- if AI components need to trigger actions, for example amending files or directing output to external systems, you apply appropriate restrictions to the possible actions (this includes external AI and non-AI fail-safes if necessary)

- decisions around user interaction are informed by AI-specific risks, for example:

  - your system provides users with usable outputs without revealing unnecessary levels of detail to a potential attacker
  - if necessary, your system provides effective guardrails around model outputs
  - if offering an API to external customers or collaborators, you apply appropriate controls that mitigate attacks on the AI system via the API
  - you integrate the most secure settings into the system by default
  - you apply least privilege principles to limit access to a system's functionality
  - you explain riskier capabilities to users and require users to opt in to use them; you communicate prohibited use cases, and, where possible, inform users of alternative solutions

### Consider security benefits and trade-offs when selecting your AI model

Your choice of AI model will involve balancing a range of requirements. This includes choice of model architecture, configuration, training data, training algorithm and hyperparameters. Your decisions are informed by your threat model, and are regularly reassessed as AI security research advances and understanding of the threat evolves.

When choosing an AI model, your considerations will likely include, but are not limited to:

- the complexity of the model you are using, that is, the chosen architecture and number of parameters; your model's chosen architecture and number of parameters will, among other factors, affect how much training data it requires and how robust it is to changes in input data when in use

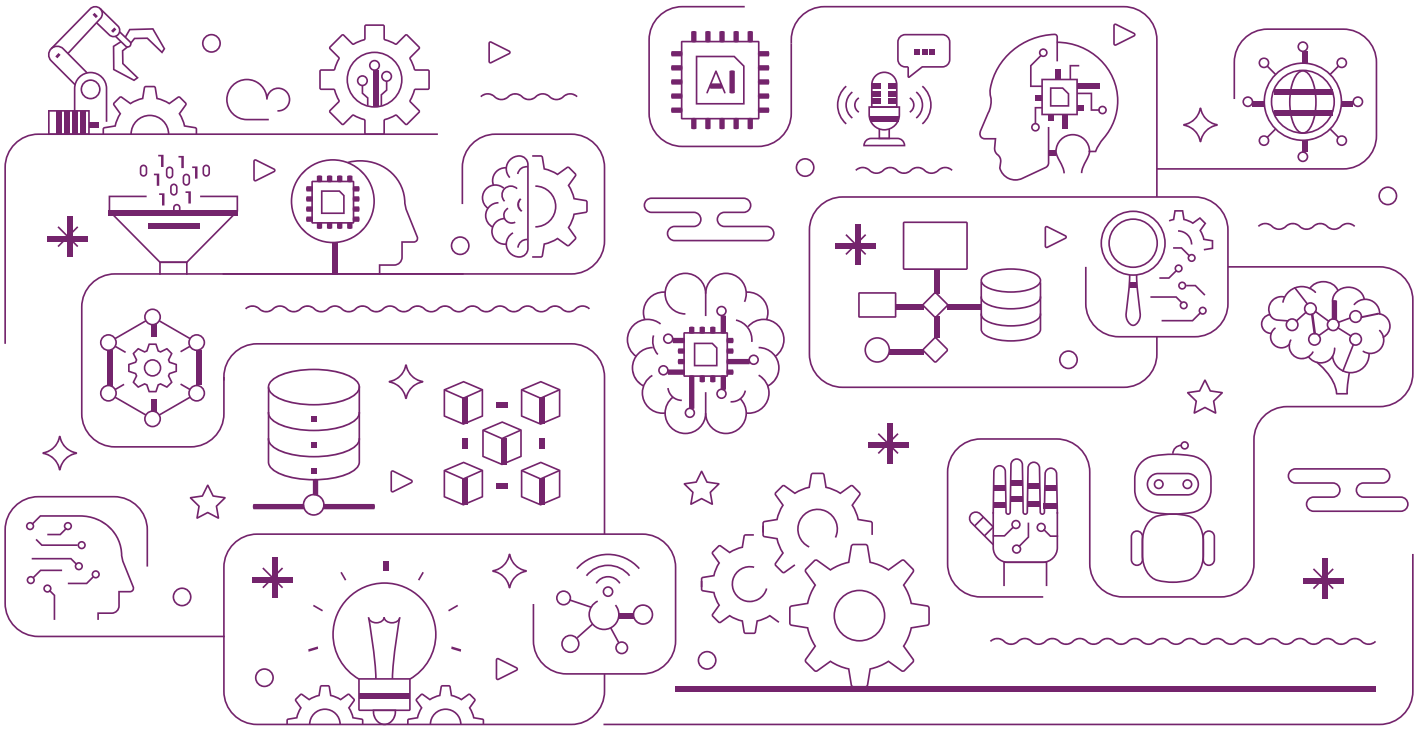
- the appropriateness of the model for your use case and/or feasibility of adapting it to your specific need (for example by fine-tuning)

- the ability to align, interpret and explain your model's outputs (for example for debugging, audit or regulatory compliance); there may be benefits to using simpler, more transparent models over large and complex ones which are more difficult to interpret

- characteristics of training dataset(s), including size, integrity, quality, sensitivity, age, relevance and diversity

- the value of using model hardening (such as adversarial training), regularisation and/or privacy-enhancing techniques
- the provenance and supply chains of components including the model or foundation model, training data and associated tools

For more information about how many of these factors impact security outcomes, refer to the NCSC's 'Principles for the Security of Machine Learning', in particular Design for security (model architecture).



# 2. Secure development

This section contains guidelines that apply to the **development** stage of the AI system development lifecycle, including supply chain security, documentation, and asset and technical debt management.

### Secure your supply chain 

You assess and monitor the security of your AI supply chains across a system's life cycle, and require suppliers to adhere to the same standards your own organisation applies to other software. If suppliers cannot adhere to your organisation's standards, you act in accordance with your existing risk management policies.

Where not produced in-house, you acquire and maintain well-secured and well-documented hardware and software components (for example, models, data, software libraries, modules, middleware, frameworks, and external APIs) from verified commercial, open source, and other third-party developers to ensure robust security in your systems.

You are ready to failover to alternate solutions for mission-critical systems, if security criteria are not met. You use resources like the NCSC's Supply Chain Guidance and frameworks such as Supply Chain Levels for Software Artifacts (SLSA)[10] for tracking attestations of the supply chain and software development life cycles.

### Identify, track and protect your assets 

You understand the value to your organisation of your AI-related assets, including models, data (including user feedback), prompts, software, documentation, logs and assessments (including information about potentially unsafe capabilities and failure modes), recognising where they represent significant investment and where access to them enables an attacker. You treat logs as sensitive data and implement controls to protect their confidentiality, integrity and availability.

You know where your assets reside and have assessed and accepted any associated risks. You have processes and tools to track, authenticate, version control and secure your assets, and can restore to a known good state in the event of compromise.

You have processes and controls in place to manage what data AI systems can access, and to manage content generated by AI according to its sensitivity (and the sensitivity of the inputs that went into generating it).

### Document your data, models and prompts 

You document the creation, operation, and life cycle management of any models, datasets and meta- or system-prompts. Your documentation includes security-relevant information such as the sources of training data (including fine-tuning data and human or other operational feedback), intended scope and limitations, guardrails, cryptographic hashes or signatures, retention time, suggested review frequency and potential failure modes. Useful structures to help do this include model cards, data cards and software bills of materials (SBOMs). The production of comprehensive documentation supports transparency and accountability[11].

### Manage your technical debt

As with any software system, you identify, track and manage your 'technical debt' throughout an AI system's life cycle (technical debt is where engineering decisions that fall short of best practices to achieve short-term results are made, at the expense of longer-term benefits). Like financial debt, technical debt is not inherently bad, but should be managed from the earliest stages of development[12]. You recognise that doing so can be more challenging in an AI context than for standard software, and that your levels of technical debt are likely to be high due to rapid development cycles and a lack of well-established protocols and interfaces. You ensure your life cycle plans (including processes to decommission AI systems) assess, acknowledge and mitigate risks to future similar systems.

# 3. Secure deployment

This section contains guidelines that apply to the **deployment** stage of the AI system development life cycle, including protecting infrastructure and models from compromise, threat or loss, developing incident management processes, and responsible release.

### Secure your infrastructure 

You apply good infrastructure security principles to the infrastructure used in every part of your system's life cycle. You apply appropriate access controls to your APIs, models and data, and to their training and processing pipelines, in research and development as well as deployment. This includes appropriate segregation of environments holding sensitive code or data. This will also help mitigate standard cyber security attacks which aim to steal a model or harm its performance.

### Protect your model continuously 

Attackers may be able to reconstruct the functionality of a model[13] or the data it was trained on[14], by accessing a model directly (by acquiring model weights) or indirectly (by querying the model via an application or service). Attackers may also tamper with models, data or prompts during or after training, rendering the output untrustworthy.

You protect the model and data from direct and indirect access, respectively, by:

- implementing standard cyber security best practices
- implementing controls on the query interface to detect and prevent attempts to access, modify, and exfiltrate confidential information

To ensure that consuming systems can validate models, you compute and share cryptographic hashes and/or signatures of model files (for example, model weights) and datasets (including checkpoints) as soon as the model is trained. As always with cryptography, good key management is essential[15].

Your approach to confidentiality risk mitigation will depend considerably on the use case and the threat model. Some applications, for example those involving very sensitive data, may require theoretical guarantees that can be difficult or expensive to apply. If appropriate, privacy-enhancing technologies (such as differential privacy or homomorphic encryption) can be used to explore or assure levels of risk associated with consumers, users and attackers having access to models and outputs.

### Develop incident management procedures 

The inevitability of security incidents affecting your AI systems is reflected in your incident response, escalation and remediation plans. Your plans reflect different scenarios and are regularly reassessed as the system and wider research evolves. You store critical company digital resources in offline backups. Responders have been trained to assess and address AI-related incidents. You provide high-quality audit logs and other security features or information to customers and users at no extra charge, to enable their incident response processes.

### Release AI responsibly

You release models, applications or systems only after subjecting them to appropriate and effective security evaluation such as benchmarking and red teaming (as well as other tests that are out of scope for these guidelines, such as safety or fairness), and you are clear to your users about known limitations or potential failure modes. Details of open-source security testing libraries are given in the further reading section at the end of this document.

### Make it easy for users to do the right things

You recognise that each new setting or configuration option is to be assessed in conjunction with the business benefit it derives, and any security risks it introduces. Ideally, the most secure setting will be integrated into the system as the only option. When configuration is necessary, the default option should be broadly secure against common threats (that is, secure by default). You apply controls to prevent the use or deployment of your system in malicious ways.

You provide users with guidance on the appropriate use of your model or system, which includes highlighting limitations and potential failure modes. You state clearly to users which aspects of security they are responsible for, and are transparent about where (and how) their data might be used, accessed or stored (for example, if it is used for model retraining, or reviewed by employees or partners).

# 4. Secure operation and maintenance

This section contains guidelines that apply to the **secure operation and maintenance** stage of the AI system development life cycle. It provides guidelines on actions particularly relevant once a system has been deployed, including logging and monitoring, update management and information sharing.

### Monitor your system's behaviour

You measure the outputs and performance of your model and system such that you can observe sudden and gradual changes in behaviour affecting security. You can account for and identify potential intrusions and compromises, as well as natural data drift.

### Monitor your system's inputs

In line with privacy and data protection requirements, you monitor and log inputs to your system (such as inference requests, queries or prompts) to enable compliance obligations, audit, investigation and remediation in the case of compromise or misuse. This could include explicit detection of out-of-distribution and/or adversarial inputs, including those that aim to exploit data preparation steps (such as cropping and resizing for images).

### Follow a secure by design approach to updates

You include automated updates by default in every product and use secure, modular update procedures to distribute them. Your update processes (including testing and evaluation regimes) reflect the fact that changes to data, models or prompts can lead to changes in system behaviour (for example, you treat major updates like new versions). You support users to evaluate and respond to model changes (for example by providing preview access and versioned APIs).

### Collect and share lessons learned

You participate in information-sharing communities, collaborating across the global ecosystem of industry, academia and governments to share best practice as appropriate. You maintain open lines of communication for feedback regarding system security, both internally and externally to your organisation, including providing consent to security researchers to research and report vulnerabilities. When needed, you escalate issues to the wider community, for example publishing bulletins responding to vulnerability disclosures, including detailed and complete common vulnerability enumeration. You take action to mitigate and remediate issues quickly and appropriately.

# Further reading

## AI development

Principles for the security of machine learning
The NCSC's detailed guidance on developing, deploying or operating a system with an ML component.

Secure by Design - Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software
Co-authored by CISA, the NCSC and other agencies, this guidance describes how manufacturers of software systems, including AI, should take steps to factor security into the design stage of product development, and ship products that come secure out of the box.

AI Security Concerns in a Nutshell
Produced by the German Federal Office for Information Security (BSI), this document provides an introduction to possible attacks on machine learning systems and potential defences against those attacks.

Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems and Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems
These documents, produced as part of the G7 Hiroshima AI Process, provide guidance for organisations developing the most advanced AI systems, including the most advanced foundation models and generative AI systems with the aim of promoting safe, secure, and trustworthy AI worldwide.

AI Verify
Singapore's AI Governance Testing Framework and Software toolkit that validates the performance of AI systems against a set of internationally recognised principles through standardised tests.

Multilayer Framework for Good Cybersecurity Practices for AI — ENISA (europa.eu)
A framework to guide National Competent Authorities and AI stakeholders on the steps they need to follow to secure their AI systems, operations and processes

ISO 5338: AI system life cycle processes (Under review)
A set of processes and associated concepts for describing the life cycle of AI systems based on machine learning and heuristic systems.

AI Cloud Service Compliance Criteria Catalogue (AIC4)
BSI's AI Cloud Service Compliance Criteria Catalogue provides AI-specific criteria, which enable evaluation of the security of an AI service across its lifecycle.

NIST IR 8269 (Draft) A Taxonomy and Terminology of Adversarial Machine Learning
A set of processes and associated concepts for describing the life cycle of AI systems based on machine learning and heuristic systems.

MITRE ATLAS
A knowledge base of adversary tactics, techniques, and case studies for machine learning (ML) systems, modelled after and linked to MITRE ATT&CK framework.

An Overview of Catastrophic AI Risks (2023)
Produced by the Center for AI Safety, this document sets out areas of risk posed by AI.

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.

Introducing Artificial Intelligence
Blog from the Australian Cyber Security Centre which provides approachable guidance on Artificial Intelligence and how to securely engage with it.

Open-source projects to help users security test AI models include:

- Adversarial Robustness Toolbox (IBM)
- CleverHans (University of Toronto)
- TextAttack (University of Virginia)
- Prompt Bench (Microsoft)
- Counterfit (Microsoft).
- AI Verify (Infocomm Media Development Authority, Singapore)

## Cyber security

CISA's Cybersecurity Performance Goals
A common set of protections that all critical infrastructure entities should implement to meaningfully reduce the likelihood and impact of known risks and adversary techniques.

NCSC CAF Framework
The Cyber Assessment Framework (CAF) provides guidance for organisations responsible for vitally important services and activities.

MITRE's Supply Chain Security Framework
A framework for evaluating suppliers and service providers within the supply chain.

## Risk management

NIST AI Risk Management Framework (AI RMF)
The AI RMF outlines how to manage socio-technical risks to individuals, organisations, and society uniquely associated with AI.

ISO 27001: Information security, cybersecurity and privacy protection
This standard provides organisations with guidance on the establishment, implementation and maintenance of an information security management system

ISO 31000: Risk management
An international standard that provides organisations with guidelines and principles for risk management within organisations

NCSC Risk Management Guidance
This guidance helps cyber security risk practitioners to better understand and manage the cyber security risks affecting their organisations.

# Notes

1. Here defined as a person, public authority, agency or other body that develops an AI system (or that has an AI system developed) and places that system on the market or puts it into service under its own name or trademark
2. For more information on secure by design, see CISA's Secure by Design web page and guidance Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software
3. As opposed to non-ML AI approaches such as rule-based systems
4. CEPS describe seven different types of AI development interaction in their publication 'Reconciling the AI Value Chain with the EU's Artificial Intelligence Act'
5. ISO/IEC 22989:2022(en) defines this as 'a functional element that constructs an AI system'
6. NIST is tasked with producing guidelines (and taking other actions) to advance the safe, secure, and trustworthy development and use of Artificial Intelligence (AI). See NIST's Responsibilities Under the October 30, 2023 Executive Order
7. More information on threat modelling is available from the OWASP Foundation
8. See MITRE ATLAS Adversarial Machine Learning 101
9. GitHub: RCE PoC for Tensorflow using a malicious Lambda layer
10. SLSA: 'Safeguarding artifact integrity across any software supply chain'
11. METI (Japanese Ministry of Economy, Trade and Industry, 2023), 'Guide of Introduction of Software Bill of Materials (SBOM) for Software Management'
12. Google research: Machine Learning: The High Interest Credit Card of Technical Debt
13. Tramèr et al 2016, Stealing Machine Learning Models via Prediction APIs
14. Boenisch, 2020, Attacks against Machine Learning Privacy (Part 1): Model Inversion Attacks with the IBM-ART Framework
15. National Cyber Security Centre, 2020, Design and build a privately hosted Public Key Infrastructure