AI Cyber Security Code of Practice

Introduction

The UK Government is taking forward a two-part intervention to address the cyber security risks to AI. This involves the development of a voluntary Code of Practice which will be used to help create a global standard in the European Telecommunication Standards Institute (ETSI) that sets baseline security requirements. We believe a Code focused specifically on the cyber security of AI is needed because AI has distinct differences to software. These include security risks from data poisoning, model obfuscation, indirect prompt injection and operational differences associated with data management. Further examples of the unique risks posed by AI systems can be found in Appendix B within the National Institute of Standards and Technology's (NIST) Risk Management Framework. The Government is also intervening in this area because software needs to be secure by design and stakeholders in the AI supply chain require clarity on what baseline security requirements they should implement to protect AI systems.

The proposed intervention was endorsed by 80% of respondents to the Department for Science, Innovation and Technology's (DSIT) Call for Views which was held from 15 May to 9 August 2024. Support for each principle in the Code ranged from 83% to 90%. This document also builds on NCSC's Guidelines for Secure AI Development which were published in November 2023 and endorsed by 19 international partners. As set out in DSIT's modular approach to cyber security codes of practice, AI stakeholders should view this document as an addendum to the Software Code of Practice.[1]

Scope

The scope of this voluntary Code of Practice is focused on AI systems. This includes systems that incorporate deep neural networks, such as generative AI. For consistency, we have used the term "AI systems" throughout the document when framing the scope of provisions and "AI security" which is considered a subset of cyber security. The Code is not designed for academics who are creating and testing AI systems only for research purposes (AI systems which are not going to be deployed).

The Code sets out cyber security requirements for the lifecycle of AI. We recognise that there is no consistent view in international frameworks on what forms the AI lifecycle. However, to help stakeholders, we have separated the principles into five phases. These are secure design, secure development, secure deployment, secure maintenance and secure end of life. We have also signposted relevant standards and publications at the start of each principle to highlight links between the various documents and the Code. This is not an exhaustive list.[2]

Implementation Guide

Following Call for Views feedback, we have created an implementation guide to support organisations with adhering to the requirements in the voluntary Code (and future standard). The guide was developed following an extensive review of software and AI standards and frameworks as well as documents published by other governments and regulators. The UK Government plan to submit the Code and Implementation Guide in ETSI so that the future

[1] Cyber security codes of practice, 31 January 2025, https://www.gov.uk/government/collections/cyber-security-codes-of-practice

[2] More information on the publication sources can be found in the "Cyber security for AI recommendations" report at https://www.gov.uk/government/publications/research-on-the-cyber-security-of-ai

standard is accompanied by a guide. The Government will update the content of the Code and Guide to mirror the future ETSI global standard and guide.

Audience

This section defines the stakeholder groups that form the AI supply chain. An indication is given for each principle on which stakeholders are primarily responsible for its implementation. Importantly, a single entity may hold multiple stakeholder roles in this voluntary Code as well as responsibilities from different regulatory regimes.[3] All stakeholders included in the table below should note that when the data used for an AI system is personal (including pseudonymized data), they may have data protection obligations and will need to consult UK data protection guidance offered by the ICO.[4] Additionally, senior leaders in an organisation also have responsibilities to help protect their staff and infrastructure as noted in DSIT’s Cyber Governance Code of Practice. Some provisions for Developers in the Code are less applicable to AI systems involving open-source models. We encourage Developers to review the Implementation Guide to confirm what requirements are specified for different types of AI systems.

Stakeholder	**Definitions**
Developers	This encompasses any type of business or organisation across any sector as well as individuals that are responsible for creating or adapting an AI model and/or system. This applies to all AI technologies, including proprietary and open-source models. For context, a business or organisation that creates an AI model and who is also responsible for embedding/deploying that model/system in their organisation would be defined in this voluntary Code to be both a Developer and a System Operator.
System Operators	This includes any type of business or organisation across any sector that has responsibility for embedding / deploying an AI model and system within their infrastructure. This applies to all AI technologies, including proprietary and open-source models. This term also includes those businesses that provide a contractual service to organisations to embed / deploy an AI model and system for business purposes.
Data Custodians	This includes any type of business, organisation or individual that controls data permissions and the integrity of data that is used for any AI model or system to function. This stakeholder group also includes those entities that set the policies for how data is used and managed for an AI model and/or system. In the context of an AI system, there could be multiple data custodians involved because some data used to create a model could come from the organisation that is deploying/embedding the system in their infrastructure and other data could be from public databases and other sources.

[3] Examples include under data protection law, when processing personal data organisations may have a role of controller and/or joint controller and/or processor, depending on their role in creating and setting up AI systems.

[4] Information Commissioner’s Office; Guidance on AI and Data Protection and Guidance on Explaining decisions made with AI

End-users	This encompasses any employee within an organisation or business and UK consumers who use an AI model and system for any purpose, including to support their work and day-to-day activities. This applies to all AI technologies and both proprietary and open-source models. This stakeholder group has been created because the voluntary Code has placed expectations on Developers, System Operators and Data custodians to help inform and protect end-users.
Affected entities	Encompasses all individuals and technologies, such as apps and autonomous systems, that are not directly affected by AI systems or decisions based on the output of AI systems. These individuals do not necessarily interact with the deployed system or application.

The table below gives examples of common cases involving different types of organisations that are relevant to this voluntary Code of Practice as well as the Software Resilience voluntary Code of Practice.

Stakeholder Groups	**Guidance**
Software vendors who also offer AI services to customers/end-users	These organisations are likely to be Developers and therefore are in scope of this Code and the Software Resilience Code of Practice.
Software vendors who use AI in their own infrastructure which has been created by an external provider	These organisations are likely to be System Operators and therefore are in scope of relevant parts of the Code and the Software Resilience Code of Practice.
Software vendors who create AI in-house and implement it within their infrastructure	These organisations are likely to be Developers and System Operators and therefore are in scope of this Code and the Software Resilience Code of Practice.
Software vendors who only use third-party AI (components) for their in-house use	These organisations are likely to be System Operators and therefore are in scope of relevant parts of the Code and the Software Resilience Code of Practice.
Organisation that creates an AI system for in-house use	These organisations are likely to be Developers and therefore are in scope of this Code.
Organisation that only uses third-party AI components	These organisations are likely to be System Operators and therefore are in scope of relevant parts of the Code.
AI Vendors	Organisations that offer or sell models and components, but do not play a role in developing or deploying them, are not likely to be in scope of this Code. These organisations are likely to be in scope

of the Software Code of Practice and Cyber Governance Code.

Terminology

We have used “shall” and “should” terminology for each provision in the voluntary Code to align with the wording used by standards development organisations. The table below sets out the definitions of these words in the context of the voluntary nature of this Code of Practice. A glossary can be found in Annex A.

Term	Definition
Shall	Indicates a requirement for the voluntary Code
Should	Indicates a recommendation for the voluntary Code
May	Indicates where something is possible, for example, that an organisation or individual is able to do something

Structure of the voluntary Code of Practice

Principle 1: Raise awareness of AI security threats and risks
Principle 2: Design your AI system for security as well as functionality and performance
Principle 3: Evaluate the threats and manage the risks to your AI system
Principle 4: Enable human responsibility for AI systems
Principle 5: Identify, track and protect your assets
Principle 6: Secure your infrastructure
Principle 7: Secure your supply chain
Principle 8: Document your data, models and prompts
Principle 9: Conduct appropriate testing and evaluation
Principle 10: Communication and processes associated with End-users and Affected Entities
Principle 11: Maintain regular security updates, patches and mitigations
Principle 12: Monitor your system’s behaviour
Principle 13: Ensure proper data and model disposal

Code of Practice Principles

Secure Design

Principle 1: Raise awareness of AI security threats and risks
Primarily applies to: System Operators, Developers, and Data Custodians

[NIST 2022, NIST 2023, ASD 2023, WEF 2024, OWASP 2024, MITRE 2024, Google 2023, ESLA 2023, Cisco 2022, Deloitte 2023, Microsoft 2022].

1.1. Organisations’ cyber security training programme shall include AI security content which shall be regularly reviewed and updated, such as if new substantial AI-related security threats emerge.

> 1.1.1 AI security training shall be tailored to the specific roles and responsibilities of staff members.

1.2. As part of an Organisation’s wider staff training programme, they shall require all staff to maintain awareness of the latest security threats and vulnerabilities that are AI-related. Where available, this awareness shall include proposed mitigations.

> 1.2.1. These updates should be communicated through multiple channels, such as security bulletins, newsletters, or internal knowledge-sharing platforms. This will ensure broad dissemination and understanding among the staff.
>
> 1.2.2 Organisations shall provide developers with training in secure coding and system design techniques specific to AI development, with a focus on preventing and mitigating security vulnerabilities in AI algorithms, models, and associated software.

Principle 2: Design your AI system for security as well as functionality and performance
Primarily applies to: System Operators and Developers

[OWASP 2024, MITRE 2024, WEF 2024, ENISA 2023, NCSC 2023, BSI1 2023, Cisco 2022, Microsoft 2022, G7 2023, HHS 2021, OpenAI2 2024, ASD 2023, ICO 2020].

2.1 As part of deciding whether to create an AI system, a System Operator and/or Developer shall conduct a thorough assessment that includes determining and documenting the business requirements and/or problem they are seeking to address, along with associated AI security risks and mitigation strategies.[5]

> 2.1.1 Where the Data Custodian is part of a Developer’s organisation, they shall be included in internal discussions when determining the requirements and data needs of an AI system.

2.2: Developers and System Operators shall ensure that AI systems are designed and implemented to withstand adversarial AI attacks, unexpected inputs and AI system failure.

2.3 To support the process of preparing data, security auditing and incident response for an AI system, Developers shall document and create an audit trail in relation to the AI system. This shall include the operation, and lifecycle management of models, datasets and prompts incorporated into the system.

2.4 If a Developer or System Operator uses an external component, they shall conduct an AI security risk assessment and due diligence process in line with their existing software development processes, that assesses AI specific risks.[6]

2.5 Data Custodians shall ensure that the intended usage of the system is appropriate to the sensitivity of the data it was trained on as well as the controls intended to ensure the security of the data.

> 2.5.1 Organisations should ensure that employees are encouraged to proactively report and identify any potential security risks in AI systems and ensure appropriate safeguards are in place.

2.6 Where the AI system will be interacting with other systems or data sources, (be they internal or external), Developers and System Operators shall ensure that the permissions granted to the AI system on other systems are only provided as required for functionality and are risk assessed.

[5] When the data processed is personal data, those risks and mitigations can be logged as part of the Data Protection Impact Assessment process (DPIA). UK General Data Protection Regulation Articles 25(1) and 25(2) set out regulatory requirements that security needs to be built at the design stage. Articles 35 and 36 set out the requirements to complete a DPIA, including where it is necessary to notify the ICO (see ICO Guidance, Do we need to consult the ICO? | ICO)

[6] See NIST Artificial Intelligence Risk Management Framework (NIST AI 100-) for examples of good practices, January 2023, https://doi.org/10.6028/NIST.AI.100-1 More detail can be found in the Implementation Guide. Additionally, see ICO Guidance: A guide to data security

2.7 If a Developer or System Operator chooses to work with an external provider, they shall undertake a due diligence assessment and should ensure that the provider is adhering to this Code of Practice.

Principle 3: Evaluate the threats and manage the risks to your AI system
Primarily applies to: Developers and System Operators

[OWASP 2024, WEF 2024, Nvidia 2023, ENISA 2023, Google 2023, G7 2023, NCSC 2023, Deloitte 2023], MITRE, OWASP, NIST Risk Taxonomy, ISO 27001]

3.1 Developers and System Operators shall analyse threats and manage security risks to their systems. Threat modelling should include regular reviews and updates and address AI-specific attacks, such as data poisoning, model inversion, and membership inference.

> 3.1.1 The threat modelling and risk management process shall be conducted to address any security risks that arise when a new setting or configuration option is implemented or updated at any stage of the AI lifecycle.
>
> 3.1.2 Developers shall manage the security risks associated with AI models that provide superfluous functionalities, where increased functionality leads to increased risk. For example, where a multi-modal model is being used but only single modality is used for system function.
>
> 3.1.3 System Operators shall apply controls to risks identified through the analysis based on a range of considerations, including the cost of implementation in line with their corporate risk tolerance.

3.2 Where AI security threats are identified that cannot be resolved by Developers, this shall be communicated to System Operators so they can threat model their systems. System Operators shall communicate this information to End-users, so they are made aware of these threats. This communication should include detailed descriptions of the risks, potential impacts, and recommended actions to address or monitor these threats.

3.3 Where an external entity has responsibility for AI security risks identified within an organisations infrastructure, System Operators should attain assurance that these parties are able to address such risks.

3.4 Developers and System Operators should continuously monitor and review their system infrastructure according to risk appetite. It is important to recognise that a higher level of risk will remain in AI systems despite the application of controls to mitigate against them.

Principle 4: Enable human responsibility for AI systems
Primarily applies to: Developers and System Operators

[OWASP 2024, MITRE 2024, BSI1 2023, Microsoft 2022]

4.1 When designing an AI system, Developers and/or System Operators should incorporate and maintain capabilities to enable human oversight.[7]

4.2 Developers should design systems to make it easy for humans to assess outputs that they are responsible for in said system (such as by ensuring that models outputs are explainable or interpretable).

4.3 Where human oversight is a risk control, Developers and/or System Operators shall design, develop, verify and maintain technical measures to reduce the risk through such oversight.

4.4 Developers should verify that the security controls specified by the Data Custodian have been built into the system.

4.5 Developers and System Operators should make End-users aware of prohibited use cases of the AI system.

[7] When the AI system processes personal data organisations should consider whether the obligations around automated decision making are applicable. Where the processing may be within scope of Article 22 of the UK GDPR, you should consult the ICO's guidance (See Automated decision-making and profiling | ICO)

Secure Development

Principle 5: Identify, track and protect your assets
Primarily applies to: Developers, System Operators and Data Custodians

[OWASP 2024, Nvidia 2023, NCSC 2023, BSI1 2023, Cisco 2022, Deloitte 2023, Amazon 2023, G7 2023, ICO 2020]

5.1 Developers, Data Custodians and System Operators shall maintain a comprehensive inventory of their assets (including their interdependencies/connectivity).

5.2 As part of broader software security practices, Developers, Data Custodians and System Operators shall have processes and tools to track, authenticate, manage version control and secure their assets due to the increased complexities of AI specific assets.

5.3 System Operators shall develop and tailor their disaster recovery plans to account for specific attacks aimed at AI systems.

> 5.3.1 System Operators should ensure that a known good state can be restored.

5.4 Developers, System Operators, Data Custodians and End-users shall protect sensitive data, such as training or test data, against unauthorised access.

> 5.4.1 Developers, Data Custodians and System Operators shall apply checks and sanitisation to data and inputs when designing the model based on their access to said data and inputs and where those data and inputs are stored. This shall be repeated when model revisions are made in response to user feedback or continuous learning.
>
> 5.4.2 Where training data or model weights could be confidential, Developers shall put proportionate protections in place.

Principle 6: Secure your infrastructure
Primarily applies to: Developers and System Operators

[OWASP 2024, MITRE 2024, WEF 2024, NCSC 2023, Microsoft 2022, ICO 2020]

6.1 Developers and System Operators shall evaluate their organisation's access control frameworks and identify appropriate measures to secure APIs, models, data, and training and processing pipelines.

6.2 If a Developer offers an API to external customers or collaborators, they shall apply controls that mitigate attacks on the AI system via the API. For example, placing limits on model access rate to limit an attacker’s ability to reverse engineer or overwhelm defences to rapidly poison a model.

6.3 Developers shall also create dedicated environments for development and model tuning activities. The dedicated environments shall be backed by technical controls to ensure separation and principle of least privilege. In the context of AI, this is particularly necessary because training data shall only be present in the training and development environments where this training data is not based on publicly available data.

6.4 Developers and System Operators shall implement and publish a clear and accessible vulnerability disclosure policy.

6.5 Developers and System Operators shall create, test and maintain an AI system incident management plan and an AI system recovery plan.

6.6 Developers and System Operators should ensure that, where they are using cloud service operators to help to deliver the capability, their contractual agreements support compliance with the above requirements.

Principle 7: Secure your supply chain
Primarily applies to: Developers, System Operators and Data Custodians

[Software Bill of Materials (SBOM) | CISA, OWASP 2024, NCSC 2023, Microsoft 2022, ASD 2023]

7.1 Developers and System Operators shall follow secure software supply chain processes for their AI model and system development.

7.2 System Operators that choose to use or adapt any models, or components, which are not well-documented or secured shall be able to justify their decision to use such models or components through documentation (for example if there was no other supplier for said component).

> 7.2.1 In this case, Developers and System Operators shall have mitigating controls and undertake a risk assessment linked to such models or components.
>
> 7.2.2 System Operators shall share this documentation with End-users in an accessible way.

7.3 Developers and System Operators shall re-run evaluations on released models that they intend on using.

7.4 System Operators shall communicate their intention to update models to End-users in an accessible way prior to models being updated.

Principle 8: Document your data, models and prompts

Primarily applies to: Developers

[OWASP 2024, WEF 2024, NCSC 2023, Cisco 2022, Microsoft 2022, ICO 2020]

8.1 Developers shall document and maintain a clear audit trail of their system design and post-deployment maintenance plans. Developers should make the documentation available to the downstream System Operators and Data Custodians.

> 8.1.1 Developers should ensure that the document 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, retention time, suggested review frequency and potential failure modes.
>
> 8.1.2 Developers shall release cryptographic hashes for model components that are made available to other stakeholders to allow them to verify the authenticity of the components.

8.2 Where training data has been sourced from publicly available sources, there is a risk that this data might have been poisoned. As discovery of poisoned data is likely to occur after training (if at all), Developers shall document how they obtained the public training data, where it came from and how that data is used in the model.

> 8.2.1. The documentation of training data should include at a minimum the source of the data, such as the URL of the scraped page, and the date/time the data was obtained. This will allow Developers to identify whether a reported data poisoning attack was in their data sets.

8.3 Developers should ensure that they have an audit log of changes to system prompts or other model configuration (including prompts) that affect the underlying working of the systems. Developers may make this available to any System Operators and End-Users that have access to the model.

Principle 9: Conduct appropriate testing and evaluation

Primarily applies to: Developers and System Operators

[OWASP 2024, WEF 2024, Nvidia 2023, NCSC 2023, ENISA 2023, Google 2023, G7 2023]

9.1 Developers shall ensure that all models, applications and systems that are released to System Operators and/or End-users have been tested as part of a security assessment process.

9.2 System Operators shall conduct testing prior to the system being deployed with support from Developers.

> 9.2.1 For security testing, System Operators and Developers should use independent security testers with technical skills relevant to their AI systems.

9.3 Developers should ensure that the findings from the testing and evaluation are shared with System Operators, to inform their own testing and evaluation.

9.4 Developers should evaluate model outputs to ensure they do not allow System Operators or End-users to reverse engineer non-public aspects of the model or the training data.

> 9.4.1 Additionally, Developers should evaluate model outputs to ensure they do not provide System Operators or End-users with unintended influence over the system.

Secure Deployment

Principle 10: Communication and processes associated with End-users and Affected Entities

As part of an organisation's wider deployment practices, they should also consider pre-deployment testing of AI systems alongside the requirements below.

10.1 System Operators shall convey to End-users in an accessible way where and how their data will be used, accessed and stored (for example, if it is used for model retraining, or reviewed by employees or partners).[8] If the Developer is an external entity, they shall provide this information to System Operators.

10.2 System Operators shall provide End-users with accessible guidance to support their use, management, integration, and configuration of AI systems. If the Developer is an external entity, they shall provide all necessary information to help System Operators.

> 10.2.1 System Operators shall include guidance on the appropriate use of the model or system, which includes highlighting limitations and potential failure modes.
>
> 10.2.2 System Operators shall proactively inform End-users of any security relevant updates and provide clear explanations in an accessible way.

10.3 Developers and System Operators should support End-users and Affected Entities during and following a cyber security incident to contain and mitigate the impacts of an incident. The process for undertaking this should be documented and agreed in contracts with End-users.

[8] When the data processed is personal data, the organisation will have transparency obligations towards the people whose data is processed (see ICO guidance for more information (Right to be informed | ICO)

Secure Maintenance

Principle 11: Maintain regular security updates, patches and mitigations
Primarily applies to: Developers and System Operators

[ICO 2020]

11.1 Developers shall provide security updates and patches, where possible, and notify System Operators of the security updates. System Operators shall deliver these updates and patches to End-users.

> 11.1.1 Developers shall have mechanisms and contingency plans to mitigate security risks, particularly in instances where updates cannot be provided for AI systems.

11.2 Developers should treat major AI system updates as though a new version of a model has been developed and therefore undertake a new security testing and evaluation process to help protect users.

11.3 Developers should support System Operators to evaluate and respond to model changes, (for example by providing preview access via beta-testing and versioned APIs).

Principle 12: Monitor your system’s behaviour
Primarily applies to: Developers and System Operators

[OWASP 2024, WEF 2024, Nvidia 2023, ENISA 2023, BSI1 2023, Cisco 2022, Deloitte 2023, G7 2023, Amazon 2023, ICO 2020]

12.1 System Operators shall log system and user actions to support security compliance, incident investigations, and vulnerability remediation.

12.2 System Operators should analyse their logs to ensure that AI models continue to produce desired outputs and to detect anomalies, security breaches, or unexpected behaviour over time (such as due to data drift or data poisoning).

12.3 System Operators and Developers should monitor internal states of their AI systems where this could better enable them to address security threats, or to enable future security analytics.

12.4 System Operators and Developers should monitor the performance of their models and system over time so that they can detect sudden or gradual changes in behaviour that could affect security.

Secure End of Life

Principle 13: Ensure proper data and model disposal
Primarily applies to: Developers and System Operators

13.1 If a Developer or System Operator decides to transfer or share ownership of training data and/or a model to another entity they shall involve Data Custodians and securely dispose of these assets. This will protect AI security issues that may transfer from one AI system instantiation to another.

13.2 If a Developer or System Operators decides to decommission a model and/or system, they shall involve Data Custodians and securely delete applicable data and configuration details.

Annex A: Glossary of terms

Adversarial AI: Describes techniques and methods that exploit vulnerabilities in the way AI systems work, for example, by introducing malicious inputs to exploit their machine learning aspect and deceive the system into producing incorrect or unintended results. These techniques are commonly used in adversarial attacks but are not a distinct type of AI system.

Adversarial Attack: An attempt to manipulate an AI model by introducing specially crafted inputs to cause the model to produce errors or unintended outcomes.

Application Programming Interface (API): A set of tools and protocols that allow different software systems to communicate and interact.

Artificial Intelligence (AI): Systems designed to perform tasks typically requiring human intelligence, such as decision-making, language understanding and pattern recognition. These systems can operate with varying levels of autonomy and adapt to their environment or data to improve performance.

Data Poisoning: A type of adversarial attack where malicious data is introduced into training datasets to compromise the AI system’s performance or behaviour.

Explainability: The ability of an AI system to provide human-understandable insights into its decision-making process.

Guardrails: Predefined constraints or rules implemented to control and limit an AI system's outputs and behaviours, ensuring safety, reliability, and alignment with ethical or operational guidelines.

Inference Attack: A privacy attack where an adversary retrieves sensitive information about the training data, or users, by analysing the outputs of an AI model.

Model Inversion: A privacy attack where an adversary infers sensitive information about the training data by analysing the AI model's outputs.

Prompt: An input provided to an AI model, often in the form of text, that directs or guides its response. Prompts can include questions, instructions, or context for the desired output.

Risk Assessment: The process of identifying, analysing and mitigating potential threats to the security or functionality of an AI system.

Sanitisation: The process of cleaning and validating data or inputs to remove errors, inconsistencies and malicious content, ensuring data integrity and security.

System Prompt: A predefined input or set of instructions provided to guide the behaviour of an AI model, often used to define its tone, rules, or operational context.

Threat Modelling: A process to identify and address potential security threats to a system during its design and development phases.

Training: The process of teaching an AI model to recognise patterns, make decisions, or generate outputs by exposing it to labelled data and adjusting its parameters to minimise errors.