



DPC: 25/30463618 DC

BSI Group Headquarters

389 Chiswick High Road London W4 4AL

Tel: +44 (0)20 8996 9000

Fax: +44 (0)20 8996 7400

www.bsigroup.com

Date: 16 April 2025

Origin: International

Latest date for receipt of comments: 09 June 2025

Project No. 2022/02212

Responsible committee: IST/33/4 Security Controls and Services

Interested committees:

Title: Draft BS ISO/IEC 27090 ISO/IEC 27090 Cybersecurity - Artificial Intelligence - Guidance for addressing security threats to artificial intelligence systems.

Please notify admin.start@bsigroup.com if you are aware of any keywords that might assist in clarifying or identifying the standard or of the content of this standard:

- i) has any issues related to 3rd party IPR, patent or copyright
- ii) affects other national standard(s)
- iii) requires additional national guidance or information

WARNING: THIS IS A DRAFT AND MUST NOT BE REGARDED OR USED AS A BRITISH STANDARD. THIS DRAFT IS NOT CURRENT BEYOND 09 June 2025

This draft is issued to allow comments from interested parties; all comments will be given consideration prior to publication. No acknowledgement will normally be sent. **See overleaf for information on the submission of comments.**

No copying is allowed, in any form, without prior written permission from BSI except as permitted under the Copyright, Designs and Patent Act 1988 or for circulation within a nominating organization for briefing purposes. Electronic circulation is limited to dissemination by e-mail within such an organization by committee members.

Further copies of this draft may be purchased from BSI Shop <http://shop.bsigroup.com> or from BSI Customer Services, Tel: +44(0) 20 8996 9001 or email cservices@bsigroup.com. British, International and foreign standards are also available from BSI Customer Services.

Information on the co-operating organizations represented on the committees referenced above may be obtained from <https://standardsdevelopment.bsigroup.com/>.

Introduction

This draft standard is based on national and international discussions. Your comments on this draft are invited and will assist in the preparation of the consequent standard.

For international standards, comments will be reviewed by the relevant UK national committee before sending the consensus UK vote and comments to the international committee, which will then decide appropriate action. If the international standard is approved, it is usual for the text to be published as a British Standard.

For national standards, comments will be reviewed by the relevant UK national committee and the resulting standards published as a British Standard.

UK Vote

Please indicate whether you consider the UK should submit a negative (with supporting technical reasons) or positive vote on this draft. Please indicate if you are aware of any reason why this draft standard should not be published as a British Standard.

Submission of Comments

- **Annotated drafts are not acceptable and will be rejected.**
- All comments should be submitted online at <https://standardsdevelopment.bsigroup.com/>. You will need to register in order to comment.

Template for comments and secretariat observations

Date: xx/xx/20xx	Document: ISO/DIS xxxx
------------------	------------------------

1	2	(3)	4	5	(6)	7
MB	Clause No./ Subclause No./Annex (e.g. 3.1)	Paragraph/ Figure/Table/Note	Type of comment	Comment (justification for change) by the MB	Proposed change by the MB	Secretariat observations on each comment submitted
	3.1	Definition 1	ed	Definition is ambiguous and needs clarifying.	Amend to read '...so that the mains connector to which no connection...'	
	6.4	Paragraph 2	te	The use of the UV photometer as an alternative cannot be supported as serious problems have been encountered in its use in the UK.	Delete reference to UV photometer.	



DRAFT International Standard

ISO/IEC DIS 27090

Cybersecurity — Artificial Intelligence — Guidance for addressing security threats and failures in artificial intelligence systems

ICS: 35.030

ISO/IEC JTC 1/SC 27

Secretariat: **DIN**

Voting begins on:
2025-04-16

Voting terminates on:
2025-07-09

This document is circulated as received from the committee secretariat.

THIS DOCUMENT IS A DRAFT CIRCULATED FOR COMMENTS AND APPROVAL. IT IS THEREFORE SUBJECT TO CHANGE AND MAY NOT BE REFERRED TO AS AN INTERNATIONAL STANDARD UNTIL PUBLISHED AS SUCH.

IN ADDITION TO THEIR EVALUATION AS BEING ACCEPTABLE FOR INDUSTRIAL, TECHNOLOGICAL, COMMERCIAL AND USER PURPOSES, DRAFT INTERNATIONAL STANDARDS MAY ON OCCASION HAVE TO BE CONSIDERED IN THE LIGHT OF THEIR POTENTIAL TO BECOME STANDARDS TO WHICH REFERENCE MAY BE MADE IN NATIONAL REGULATIONS.

RECIPIENTS OF THIS DRAFT ARE INVITED TO SUBMIT, WITH THEIR COMMENTS, NOTIFICATION OF ANY RELEVANT PATENT RIGHTS OF WHICH THEY ARE AWARE AND TO PROVIDE SUPPORTING DOCUMENTATION.

Reference number
ISO/IEC DIS 27090:2025(en)

Full document details: 25/30463618 DC

© ISO/IEC 2025  **Documents**
Document Sharing Hub



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2025

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword	v
Introduction	vi
1 Scope	1
2 Normative references	1
3 Terms and definitions	1
4 Abbreviated terms	3
5 Application of information security	4
5.1 General	4
5.2 Zero trust principles	5
5.3 AI governance program	5
5.4 Applying software engineering best practices to AI	5
5.5 Supply chain management	6
5.6 Reduce impact by data limitation	6
5.6.1 General	6
5.6.2 Data minimization	7
5.6.3 Ensure allowed data	7
5.6.4 Minimal retention	7
5.6.5 Obfuscate data	7
5.6.6 Minimize open technical details	7
5.7 Controlling the effects of unwanted model behaviour during the operation stage	8
5.7.1 General	8
5.7.2 Oversight	8
5.7.3 Least model privilege	9
5.7.4 Transparency	9
5.7.5 Continuous validation	9
5.7.6 Explainability	9
5.7.7 Unwanted bias testing	10
5.8 Threat modelling	10
5.9 Red teaming	10
6 Threats to AI systems	10
6.1 General	10
6.2 Data poisoning attack	11
6.2.1 Description	11
6.2.2 Failures resulting from data poisoning attacks	12
6.2.3 Detection of data poisoning attacks	13
6.2.4 Mitigation of data poisoning attacks	13
6.3 Evasion attack	15
6.3.1 Description	15
6.3.2 Failures resulting from evasion attacks	16
6.3.3 Detection of evasion attacks	16
6.3.4 Mitigation of evasion attacks	16
6.4 Membership inference	18
6.4.1 Description	18
6.4.2 Failures resulting from membership inference	18
6.4.3 Detection of membership inference	19
6.4.4 Mitigation of membership inference	19
6.5 Model exfiltration	19
6.5.1 Description	19
6.5.2 Failures resulting from model exfiltration	20
6.5.3 Detection of model exfiltration	20
6.5.4 Mitigation of model exfiltration	21
6.6 Model inversion	21
6.6.1 Description	21

6.6.2	Failures resulting from model inversion	21
6.6.3	Detection of model inversion	22
6.6.4	Mitigation of model inversion	22
6.7	Direct model poisoning	22
6.7.1	Description	22
6.7.2	Failures resulting from direct model poisoning	23
6.7.3	Detection of direct model poisoning	23
6.7.4	Mitigation of direct model poisoning	23
6.8	Direct model theft	23
6.8.1	Description	23
6.8.2	Failures resulting from direct model theft	23
6.8.3	Detection of direct model theft	23
6.8.4	Mitigations of direct model theft	23
6.9	Direct training data leak	24
6.9.1	Description	24
6.9.2	Failures resulting from training data leak	24
6.9.3	Detection of training data leak	24
6.9.4	Mitigations of training data leak	24
6.10	Model input leak	26
6.10.1	Description	26
6.10.2	Failures resulting from model input leak	26
6.10.3	Detection of model input leak	26
6.10.4	Mitigations of model input leak	26
6.11	Sensitive model output	26
6.11.1	Description	26
6.11.2	Failures resulting from sensitive model output	26
6.11.3	Detection of sensitive model output	27
6.11.4	Mitigations of sensitive model output	27
6.12	Prompt injection	27
6.12.1	Description	27
6.12.2	Failures resulting from prompt injection	27
6.12.3	Detection of prompt injection	28
6.12.4	Mitigation of prompt injection	28
6.13	Output contains injection attacks	28
6.13.1	Description	28
6.13.2	Failures resulting from output contains injection attacks	28
6.13.3	Detection of output contains injection attacks	28
6.13.4	Mitigations of output contains injection attacks	28
7	Mitigations and their interactions with threats and other mitigations	29
7.1	Overview	29
7.2	Conflicting interactions of mitigations	29
7.3	Continuity of mitigations across the AI life cycle	29
7.4	Model and Mitigation Deterioration Over Time	29
7.5	Logging and monitoring	30
7.6	Development environment protection	30
7.7	Malicious input detection and filtering	31
7.7.1	Detecting odd input	31
7.7.2	Detecting adversarial input	31
7.8	Throttling model use	31
7.9	Obscuring confidence	32
7.10	Attaining a sufficiently small model	32
	Annex A (informative) Mapping attacks to AI system life cycle and to assets	33
	Annex B (informative) AI-specific versions of conventional attacks	40
	Bibliography	42

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee JTC 1, Information technology, Subcommittee SC 27, Information security, cybersecurity and privacy protection.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

Mitigating the risks of security threats to AI systems in a timely manner can help to improve the trustworthiness of organizations that use or develop AI systems as well as improve the trust in AI systems. As technology continues to advance and innovate at a fast pace, emerging and new security vulnerabilities created in AI systems will require continuous evaluation of security objectives.

AI systems are subject to all of the traditional cybersecurity vulnerabilities, exploits and attacks. They are also subject to some new threats, such as those involving data and manipulated input to AI models. However, these are an additional concern and not a replacement for conventional cybersecurity concerns. As the overall cybersecurity threat landscape evolves, the need to protect AI systems against non-AI system specific threats will be an ongoing pursuit.

AI systems have become more widespread as organizations embrace digital transformation, resulting in an increasing likelihood of security attacks on these systems. Furthermore, as a result of the disparate and widespread use of AI systems, and especially within safety critical contexts, the result of security attacks can be severe and in some cases lead to physical or mental repercussions to individuals.

Due to the way AI systems are developed, and their strong dependency on training data, which can include user data or unvetted data, these systems can have additional vulnerabilities compared to conventional information processing systems. This document provides guidance to organizations who are choosing to develop or deploy AI systems, providing awareness of the types of security threats such systems can face, and helping organizations understand the consequences of such threats and the possible techniques to detect and mitigate such threats. Such awareness and understanding of this new threat landscape is imperative for securing AI systems against adversarial attacks.

This document describes various security threats and the exploitation of vulnerabilities that can happen when using AI systems. The objective is to identify recommended measures to improve the security of AI systems by identifying threats and the consequences of these threats. The identification of threats will help users and developers of AI systems better understand security issues related to AI systems.

It can be challenging to detect some threats to AI systems and thus, developing controls without effective threat detection would not be useful or adequate. This document can help identify where security threats can be effectively addressed.

Security threats and vulnerabilities with a lack of controls in AI systems can lead to failures, impacting both individuals and businesses.

Attacks such as data poisoning can cause undesirable outcomes by injecting malicious or unwanted data into training data. Where continuous learning occurs, ongoing training during deployment can be exploited through corrupted data during operation. In general, it is recommended that the established best security practices in ISO/IEC 27001^[1] are adhered to. However, some additional controls may be pertinent to AI or machine learning (ML)-enabled work, data, and control flows. This document will focus on those aspects, while encouraging and assuming the reader is familiar with the established requirements. Existing measures such as access control can be used to ensure that only authorized users can access training data. However, further measures are required to ensure that the data being used will provide the expected performance. While detection of data poisoning is a necessary component of such measures, it is often difficult to perform such threat detection. This document will provide the necessary insights to apply both established and new controls, for example by explaining what the assets are, and what aspects need protecting.

Other attacks such as model exfiltration can be mitigated by conventional information security measures such as rate limiting. Models and intellectual property can be stolen even when effective conventional information security measures are in place, since a model exfiltration attack is established through legitimate querying of the model. Threat detection is a challenge but an essential component of effectively mitigating model exfiltration. The measures used to mitigate AI threats can use a defence in depth strategy that leverages multiple security measures to protect an organization's assets.

If a choice is made to use unvetted or unverified data for training, risks can increase related to data poisoning or attacks that can adjust ML data or control flows.

Identification of patent holders: the following text shall be included if patent rights have been identified.

The International Organization for Standardization (ISO) [and/or] International Electrotechnical Commission (IEC) draw[s] attention to the fact that it is claimed that compliance with this document may involve the use of a patent.

ISO [and/or] IEC take[s] no position concerning the evidence, validity and scope of this patent right.

The holder of this patent right has assured ISO [and/or] IEC that he/she is willing to negotiate licences under reasonable and non-discriminatory terms and conditions with applicants throughout the world. In this respect, the statement of the holder of this patent right is registered with ISO [and/or] IEC. Information may be obtained from the patent database available at www.iso.org/patents.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights other than those in the patent database. ISO [and/or] IEC shall not be held responsible for identifying any or all such patent rights.

Cybersecurity — Artificial Intelligence — Guidance for addressing security threats and failures in artificial intelligence systems

1 Scope

This document provides guidance for organizations to address security threats and failures specific to artificial intelligence (AI) systems. The guidance in this document aims to provide information to organizations to help them better understand the consequences of security threats specific to AI systems, throughout their life cycle, and descriptions of how to detect and mitigate such threats.

This document is applicable to all types and sizes of organizations, including public and private companies, government entities, and not-for-profit organizations, that develop or use AI systems.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 22989:2022, *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*

ISO/IEC 23053:2022, *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 22989:2022, ISO/IEC 23053:2022 and the following apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

- ISO Online browsing platform: available at <https://www.iso.org/obp>
- IEC Electropedia: available at <https://www.electropedia.org/>

3.1

availability

property of being accessible and usable on demand by an authorized entity

[SOURCE: ISO/IEC 27000:2018, 3.7]

3.2

confidentiality

property that information is not made available or disclosed to unauthorized individuals, entities, or processes

[SOURCE: ISO/IEC 27000:2018, 3.10]

3.3

integrity

property of accuracy and completeness

[SOURCE: ISO/IEC 27000:2018, 3.36]

3.4

threat

potential cause of an unwanted incident, which can result in harm to a system or organization

[SOURCE: ISO/IEC 27000:2018, 3.74]

3.5

attack

attempt to destroy, expose, alter, disable, steal or gain unauthorized access to or make unauthorized use of an asset

[SOURCE: ISO/IEC 27000:2018(en), 3.2]

3.6

failure

termination of the ability of a product to perform a required function or its inability to perform within previously specified limits

Note 1 to entry: Adapted from IEEE 610.12-1990.

[SOURCE: ISO/IEC 25040:2011(en), 4.26]

3.7

mitigation

limitation of any negative consequence of a particular incident

[SOURCE: ISO 22300:2021(en), 3.1.154]

3.8

vulnerability

weakness of an asset or control that can be exploited by one or more threats

[SOURCE: ISO/IEC 27000:2018, 3.77]

3.9

zero-knowledge attack

attack on a system that assumes no prior knowledge about the internal information of the system under attack

Note 1 to entry: The adversary can use context or historical information to infer model vulnerability. The attacker can probe the system to inform system vulnerabilities.

3.10

perfect-knowledge attack

attack on a system that assumes full understanding about the system, including internal information

Note 1 to entry: In the case of an AI system, full knowledge about the internal architecture, variables or current state of the model can be known by the attacker.

Note 2 to entry: This type of attack does not assume knowledge of credentials.

3.11

partial-knowledge attack

attack carried out by an attacker who has full understanding of the system inputs and outputs, but only a limited understanding of its internal design and operations

Note 1 to entry: This type of attack does not assume knowledge of credentials.

3.12

poisoning attack

data poisoning attack

attack where the attacker manipulates data (potentially including parameters) to affect the AI system's behaviour

3.13

evasion attack

attack where the attacker is able to avoid correct behaviour of an AI system on specific input or classes of inputs

3.14

membership inference attack

membership inference

attack where the attacker is able to attribute training data membership

3.15

model exfiltration attack

model exfiltration

model theft

model stealing attack

model extraction attack

attack where the attacker is able to copy the model (or sufficient parts thereof) either through direct access, or through (potentially repeated) use of attacker supplied inputs and observing corresponding outputs of the AI system

3.16

model inversion attack

model inversion

attack where the attacker is able to retrieve the original data (or sufficient parts thereof) that was used to generate the model through (potentially repeated) use of attacker supplied inputs and observing corresponding outputs of the AI system

4 Abbreviated terms

AI	artificial intelligence
API	application programming interface
BPDA	bandit-based derivative-free policy optimization algorithm
DoS	denial of service
DP	differential privacy
EOT	expectation over transformation
GAN	generative adversarial network
IDBT	input distortion based techniques
JEA	just enough access
JIT	just-in-time
LLM	large language model
ML	machine learning
NIST	National Institute of Standards and Technology

OOD	out of distribution
PII	personally identifiable information
RAG	retrieval augmented generation
RT	random transformations

5 Application of information security

5.1 General

Attacks on AI systems can occur throughout the AI system life cycle described in ISO/IEC 22989 and ISO/IEC 5338.^[2] The former provides AI concepts and terminology, the latter describes AI system life cycle processes. Attacks can be classified as attacks during the development and training of AI systems or during the inference or operation of AI systems, or both. AI systems are susceptible to attacks considered conventional to all types of information systems. [Annex B](#) includes examples of such attacks that can have particular considerations specific to AI system, including AI system specific consequences resulting from such attacks. Conventional security attacks on information processing systems can still affect AI systems and thus, information security controls, such as those described in ISO/IEC 27002,^[3] should be implemented, based on the needs of the organization, the information security systems, and the risks identified as a result of carrying out an information security risk assessment. The organization should consider the views and key concepts outlined in ISO/IEC 27002:2022^[3] Clause 4.2, including: control types, information security properties, cybersecurity concepts, operational capabilities, and security domains. Information security is commonly defined using the CIA triad: confidentiality, integrity, and availability. Confidentiality ensures that information is accessible only to authorized individuals and entities. Integrity ensures that information is accurate and complete. Availability ensures that information is readily available to authorized individuals when they need it. AI systems should be considered as software components that are usually integrated into another system, applications or data files rather than as a standalone AI or ML model. Therefore, security measures applied to software and to applications apply as well to AI systems in all life cycle stages, such as design, deployment, or using an AI system in operation. Validation inputs for multi-modal or foundational models, when it can be either for example text or video, can necessitate different types of controls.

The implementation of best practices to protect against information security incidents is paramount for the protection of AI systems. In some cases, additional, or adaptations of security measures can be required to adequately address the unique threats to AI systems. Such security measures can be incorporated as part of risk management of the AI system, such as described in ISO/IEC 23894^[4] or ^[5] These controls can be used to mitigate identified risks as part of a responsible AI framework for developing or using AI systems such as the AI management system defined in ISO/IEC 42001.^[6] Diligence should be used when navigating management system standards for both security (ISO/IEC 27001^[1]) and AI. Examples of existing measures that are useful in the detection and mitigation of security attacks on AI systems include integrity checks on datasets and denial of service mitigation. Logging and monitoring events related to the operation of the AI system can aid in the analysis of undesirable and anomalous system and user behaviour. Verification that the AI model has not been tampered with is also important to ensure validity of operational results and avoid unintended bias. Ensuring datasets are accurately labelled and classified can help to appropriately handle and secure input datasets. Such datasets should be appropriately secured with access controls and data loss prevention techniques applied as necessary. The AI system, like other information systems, should have processes such as threat modelling, security testing and review and vulnerability fixes applied.

Security controls described in ISO/IEC 27002^[3] (if applicable) such as those related to secure system architecture and engineering principles as described in ISO/IEC 27002:2022,^[3] 8.27, should be implemented throughout the AI system life cycle. [Annex A](#) provides further information mapping attacks on AI systems across the AI system life cycle and to assets. Security threats including but not limited to those specified in [Claus 6](#) of this document should be taken into account throughout the AI system life cycle as per controls A.6.2.3 and A.6.2.6 from ISO/IEC 42001:2023,^[6] Table A.1. The secure software development framework (SSDF),^[7] published by the national institute of standards and technology (NIST) can be used for processes to better secure software, including AI systems^[8].

The reader should consider the threats according to the system they are dealing with using a risk-based approach. The information described here are examples and should not be considered as a reference.

5.2 Zero trust principles

As part of an holistic approach to secure system engineering throughout the AI system life cycle, zero trust principles can effectively protect AI systems. For each of the attacks described in this document, the zero trust principles that are effective in helping to protect AI systems are included in the mitigations measures associated for each attack described in [Clause 6](#).

ISO/IEC 27002,^[3] 8.27 includes zero trust principles to help ensure information systems are securely designed, implemented and operated within the development life cycle. These principles can be summarized as:

- a) assume the organization's information systems are already breached;
- b) employ a "never trust and always verify" approach for access to information systems;
- c) ensure requests to information systems use end-to-end encryption;
- d) verify each request to an information system as if it originated externally;
- e) use least privilege access control, including access to interacting with the model;
- f) always authenticate requesters and always validate authorization requests.

5.3 AI governance program

To ensure that AI systems and their use within an organisation are subject to proper governance (including security), a program to govern the use of AI should be installed and executed. This means taking responsibility for AI as an organization, by keeping an inventory of AI systems and their use, perform risk analysis on them, and managing those risks.

For guidance on implementation, see ISO/IEC 42001 AI management system.^[6] ISO/IEC 42001^[6] is about extending your risk management system - it focuses on governance. Additionally, ISO/IEC 5338^[2] is about extending your software lifecycle practices. ISO/IEC 42001^[6] is a management system for the governance of responsible AI in an organization, similar to how ISO/IEC 27001^[1] is a management system for information security. ISO/IEC 42001^[6] doesn't go into the lifecycle processes. It for example does not discuss how to train models, how to do data lineage, continuous validation, versioning of AI models, project planning challenges, and how and when exactly sensitive data is used in engineering.

5.4 Applying software engineering best practices to AI

In order to facilitate secure software development (including design), software engineering best practices should be applied to AI development: general best practices and security-related practices. This includes automated testing, code quality control, code review, coding guidelines, security architecture, documentation, versioning, threat modelling, and penetration testing.

A best practice is to mix data scientist profiles with software engineering profiles in teams, as software engineers typically need to learn more about data science, and data scientists generally need to learn more about creating future-proof, secure, maintainable, and testable code.

Apart from conventional software best practices, there are important AI-specific engineering practices, including for example training data provenance and lineage, model traceability and AI-specific testing such as continuous validation, testing for model staleness and concept drift.

Other applicable standards include:

- ISO/IEC 5338^[2] - AI lifecycle covers the complete software development lifecycle for AI, by extending the existing ISO/IEC/IEEE 12207 standard on software lifecycle: defining several new processes and discussing AI-specific particularities for existing processes;

- ISO/IEC 27002:2022^[3] control 5.37 Documented operating procedures, as a small part of the development lifecycle;
- ISO/IEC 27002:2022^[3] control 8.25 Secure development lifecycle;
- ISO/IEC 27115^[9] (Cybersecurity evaluation of complex systems).

5.5 Supply chain management

Supply chain management intends to minimize the security (and other) risk from externally obtained elements. In software engineering these elements are source code or software components (e.g. open source). The particularities for AI are:

- a) supplied elements can include data and models;
- b) many of the software components are executed in the development environment instead of just in the operation stage (the runtime of the application);
- c) the development environment has AI-specific assets that are under attack threat, such as training data and model parameters.

Security risks in obtained data or models can arise from accidental mistakes or from manipulations – just like with obtained source code or software components, data or models may involve multiple suppliers. For example: a model is trained by one vendor and then fine-tuned by another vendor. Or, an AI system contains multiple models, one is a model that has been fine-tuned with data from source X, using a base model from vendor A that claims data is used from sources Y and Z, where the data from source Z was labelled by vendor B. Because of this supply chain complexity, data and model provenance is a helpful activity. The Software Bill Of Materials (SBOM) becomes the AI Bill Of Materials (AIBOM) or Model Bill of Material (MBOM).

Standard supply chain management includes provenance and pedigree, verifying signatures, using package repositories, frequent patching, and using dependency verification tools.

As said, in AI many of the software components are executed in the development environment, instead of just in the operation stage. Data engineering and model engineering involve operations on data and models for which often external components are used (e.g. tools such as Notebooks, or other MLOps applications). Because AI development has AI-specific assets such as the data and model parameters, these components pose a new threat.

For more information, refer to:

- a) ISO/IEC 27002:2022^[3] Controls 5.19, 5.20, 5.21, 5.22, 5.23, 8.30. Gap: cover this control fully, with said particularities (e.g. lacking controls on data provenance);
- b) ISO/IEC CD 5181^[10] (Data provenance) covers the data provenance aspect to complete the coverage together with the ISO/IEC 27002 controls – provided that the provenance concerns all sensitive data and is not limited to personal data;
- c) ISO/IEC 42001:2023^[6] (AI management) briefly mentions data provenance and refers to WD ISO/IEC CD 5181:2023^[10] in section B.7.5.

5.6 Reduce impact by data limitation

5.6.1 General

The impact of security threats on confidentiality and integrity can be reduced by limiting the data attack surface, meaning that the amount and the variety of data is reduced as much as possible, as well as the duration for which it is kept. This section describes several practices to apply this limitation.

5.6.2 Data minimization

Data fields or records (e.g. from a training dataset) that are unnecessary for the application and for sufficient accuracy of the AI model can be removed or anonymized, in order to prevent potential data leaks or manipulation.

A method to determine which fields or records can be removed is to statistically analyze which data elements do not play a role in model performance. Additional data minimization can be achieved through dimensional reduction and feature selection.

5.6.3 Ensure allowed data

Data (e.g. from a training set) that is prohibited for the intended purpose can be removed. This is particularly important if consent was not given and the data contains personal information collected for a different purpose. See ISO/IEC 23894^[4] (AI risk management) Section A.8 Privacy.

Minimisation of certain personal data can be required by applicable policies, laws or regulations. If this data plays a role in the accuracy of the model, these rules need to be weighed against the purpose and benefits of that accuracy.

5.6.4 Minimal retention

Once data it is no longer needed, it can be removed or anonymized - or when legally required (e.g. due to privacy laws).

Limiting the retention period of data can be seen as a special form of data minimization. Privacy regulations typically require personal data to be removed when it is no longer needed for the purpose for which it was collected. Sometimes exceptions need to be made because of other rules (e.g. to keep a record of proof). Apart from these regulations, it is a general best practice to remove any sensitive data when it is no longer of use, to reduce the impact of a data leak.

5.6.5 Obfuscate data

A degree of obfuscation of sensitive data should be attained where possible.

When this is done for personal data, it is referred to as differential privacy which is a framework for formalizing privacy in statistical and data analysis, ensuring that the privacy of individual data entries in a database is protected. The key idea is to make it possible to learn about the population as a whole while providing strong guarantees that the presence or absence of any single individual in the dataset does not significantly affect the outcome of any analysis. This is often achieved by adding a controlled amount of random noise to the results of queries on the database. This noise is carefully calibrated to mask the contribution of individual data points, which means that the output of a data analysis (or query) should be essentially the same, whether any individual's data is included in the dataset or not. In other words by observing the output, one should not be able to infer whether any specific individual's data was used in the computation.

5.6.6 Minimize open technical details

Access to technical details that can help attackers should be minimized. This reduces the information available to attackers, which can assist them in selecting and tailoring their attacks, thereby lowering the probability of a successful attack.

Minimizing and protecting technical details can be achieved by incorporating such details as an asset into information security management. This will ensure proper asset management, data classification, awareness education, policy, and inclusion in risk analysis.

NOTE this control needs to be weighed against transparency considerations that require to be more open about technical aspects of the model. The key is to minimize information that can help attackers while being transparent.