SLOVENSKI STANDARD
oSIST prEN 18282:2026

01-julij-2026

Umetna inteligenca - Specifikacije za kibernetsko varnost za sisteme UI

Artificial intelligence - Cybersecurity specifications for AI Systems

Künstliche Intelligenz - Cybersicherheitsspezifikationen für KI-Systeme

Intelligence artificielle - Spécifications en matière de cybersécurité pour les systèmes d'IA

Ta slovenski standard je istoveten z: **prEN 18282**

ICS:

35.030	Informacijska varnost	IT Security
35.240.01	Uporabniške rešitve informacijske tehnike in tehnologije na splošno	Application of information technology in general

oSIST prEN 18282:2026 **en,fr,de**

EUROPEAN STANDARD

NORME EUROPÉENNE

EUROPÄISCHE NORM

DRAFT
prEN 18282

May 2026

ICS 35.030; 35.240.01

English version

Artificial intelligence - Cybersecurity specifications for AI Systems

Intelligence artificielle - Spécifications en matière de cybersécurité pour les systèmes d'IA

Künstliche Intelligenz - Cybersicherheitsspezifikationen für KI-Systeme

This draft European Standard is submitted to CEN members for enquiry. It has been drawn up by the Technical Committee CEN/CLC/JTC 21.

If this draft becomes a European Standard, CEN and CENELEC members are bound to comply with the CEN/CENELEC Internal Regulations which stipulate the conditions for giving this European Standard the status of a national standard without any alteration.

This draft European Standard was established by CEN and CENELEC in three official versions (English, French, German). A version in any other language made by translation under the responsibility of a CEN and CENELEC member into its own language and notified to the CEN-CENELEC Management Centre has the same status as the official versions.

CEN and CENELEC members are the national standards bodies and national electrotechnical committees of Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Republic of North Macedonia, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Türkiye and United Kingdom.

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.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.

Warning : This document is not a European Standard. It is distributed for review and comments. It is subject to change without notice and shall not be referred to as a European Standard.

CEN-CENELEC Management Centre:
Rue de la Science 23, B-1040 Brussels

Ref. No. prEN 18282:2026 E

Contents

Page

European foreword

This document (prEN 18282:2026) has been prepared by Technical Committee CEN/CENELEC JTC21 "Artificial intelligence", Working Group 5 "Joint standardization on cybersecurity for AI systems", the secretariat of which is held by DS.

This document is currently submitted to the CEN Enquiry.

This document has been prepared under a standardization request addressed to CEN/CENELEC by the European Commission. The Standing Committee of the EFTA States subsequently approves these requests for its Member States.

For the relationship with EU Legislation, see informative Annex ZA which is an integral part of this document.

Introduction

The increasing use of artificial intelligence (AI) systems introduces new and evolving cybersecurity challenges. These challenges arise from AI-specific attack surfaces, complex data pipelines, learning mechanisms, and dependencies on software, hardware, and services. In addition to conventional cybersecurity risks, AI systems can be exposed to threats such as data poisoning, model poisoning, adversarial manipulation, confidentiality attacks, and exploitation of model flaws. Such threats can compromise the accuracy, robustness, security, and trustworthiness of AI systems and can lead to harm to safety, health, or fundamental rights.

This document specifies cybersecurity requirements and measures that address vulnerabilities, threats and cybersecurity risks that are specific to AI systems.

The objective of this document is to support the implementation of cybersecurity requirements for AI systems in accordance with Article 15(5) of Regulation (EU) 2024/1689 (AI Act).

This document focuses on cybersecurity aspects that arise from the characteristics of AI systems, including AI models, training and inference data, and their interaction with the operational environment.

This document does not establish a separate risk management system but provides methods to identify AI-specific cybersecurity circumstances, vulnerabilities, threats and risks that can be considered within the overall AI system risk management process.

In addition to the AI-specific cybersecurity aspects addressed in this document, the cybersecurity of AI systems also depends on the security of the underlying ICT infrastructure, software components and operational environments. These aspects are typically addressed by applicable ICT cybersecurity standards and established cybersecurity practices.

The effectiveness of AI cybersecurity measures depends on their implementation and the relevant circumstances, including operational constraints, properties of the training and inference data of the used model and architecture, system scale, and the real-world environment in which the AI system operates. It follows an approach, recognising that AI-specific cybersecurity risks cannot be fully eliminated. This document therefore focuses on achieving an appropriate level of cybersecurity and resilience against attempts by unauthorized third parties to alter the use, outputs or performance of the AI system, taking into account the intended purpose, operational context and foreseeable misuse.

This document focuses exclusively on AI-specific cybersecurity aspects. It is intended to be used in conjunction with other harmonized standards supporting the EU AI Act.

Conventional cybersecurity requirements for ICT infrastructure supporting AI systems remain applicable but are outside the scope of this document.

This document needs to be interpreted and applied in accordance with the following principles:

Principle 1: This document applies to AI systems that qualify as high-risk AI systems under the EU AI Act. It can also apply to AI systems that are not classified as high-risk under Regulation (EU) 2024/1689.

Principle 2: This document addresses AI-specific cybersecurity risks and complements other technical, organizational, and legal measures required for AI systems including conventional cybersecurity requirements. It does not replace non-cybersecurity requirements such as those related to data governance, human oversight, or fundamental rights protection.

Principle 3: AI cybersecurity techniques, tools, and practices are evolving and cannot be equally mature or available for all AI system types and contexts. Where specific techniques are not reasonably attainable, the provider can apply alternative measures, provided these measures can be justified by the provider in the technical documentation.

Principle 4: It is not possible to eliminate all cybersecurity risks. This document describes measures aimed at achieving appropriate cybersecurity resilience and does not provide guarantees against all

possible attacks. Its objective is to support proportionate, evidence-based mitigation of AI-specific cybersecurity risks in line with the state of the art.

1 Scope

This document addresses organizational and technical solutions aimed at ensuring the cybersecurity of high-risk AI systems over the life cycle, appropriate to the relevant circumstances and the risks. The technical solutions to address AI-specific vulnerabilities include, where appropriate, measures to prevent, detect, respond to, resolve and control for attacks trying to manipulate the training dataset (data poisoning), or pre-trained components used in training (model poisoning), inputs designed to cause the model to make a mistake (adversarial examples or model evasion), confidentiality attacks or model flaws. This document provides objective criteria to enable decisions on whether a given technical or organizational solution adequately achieves a given vulnerability-related goal.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

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

— ISO Online browsing platform: available at http://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

3.1 Terms from EU AI Act and other EU regulations

3.1.1
AI system
machine-based system that is designed to operate with varying levels of autonomy and that can exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments

Note 1 to entry: The verb “can” represents a possibility, not all AI systems that fit the above definition have this ability to adapt after deployment.

[SOURCE: EU AI Act 2024/1689 (Article 3(1)), modified - "a" has been removed, “may” replaced with “can” based on the use of verbs in standards. Added Note 1 to entry.]

3.1.2
cybersecurity for AI systems
activities necessary to ensure that *AI system* (3.1.1) is resilient against attempts to alter its use, behaviour, performance or compromise its security properties by malicious third parties exploiting the system’s vulnerabilities

[SOURCE: EU AI Act 2024/1689 Recital 76]

3.1.3
cybersecurity risk for AI systems
potential for loss or disruption caused by an incident to an *AI system* (3.1.1) and expressed as a combination of the magnitude of such loss or disruption and the *probability* (3.2.25) of the occurrence of the incident

Note 1 to entry: Loss or disruption affecting an *AI system* (3.1.1) can result from impacts on the *confidentiality* (3.2.6), *integrity* (3.2.24) or *availability* (3.2.4) of the *AI system* (3.1.1) or its assets.

Note 2 to entry: Cybersecurity risks for AI systems can arise when cybersecurity threats exploit vulnerabilities in the *AI system* (3.1.1) or its supporting infrastructure.

Note 3 to entry: *cybersecurity risk for AI systems* (3.1.3) is associated with the potential that threats will exploit vulnerabilities of an AI system's asset or group of assets and thereby cause harm to an organization. This is based on the Note 6 to entry of "*risk*" in EN ISO/IEC 27000 [1].

[SOURCE: EU CRA Article 3(37): modified with added text to set the scope for the incident and the risk to be limited to an AI system. Notes 1, 2 and 3 to entry added.]

3.1.4
deployer
natural or legal person, public authority, agency or other body using an *AI system* (3.1.1) under its authority except where the *AI system* (3.1.1) is used in the course of a personal non-professional activity

[SOURCE: EU AI Act 2024/1689 (Article 3(4)), modified removed "a"]

3.1.5
distributor
natural or legal person in the supply chain, other than the *provider* (3.1.9) or the importer, that makes an *AI system* (3.1.1) available on the Union market

[SOURCE: EU AI Act 2024/1689 (Article 3(6))]

3.1.6
fundamental rights
basic rights and freedoms that apply to everyone within the EU, ensuring dignity, fairness, respect, and equality

Note 1 to entry: These rights are enshrined in the Charter of Fundamental Rights of the European Union. The Charter outlines these rights, which are legally binding on EU institutions and member states when implementing EU law.

Note 2 to entry: The concept of *fundamental rights* (3.1.6) is defined in Regulation (EU) 2024/1689 and further addressed in the horizontal harmonized standards developed under CEN-CLC/JTC 21.

[SOURCE: EU charter of fundamental rights.]

3.1.7
input data
data provided to or directly acquired by an *AI system* (3.1.1) on the basis of which the system produces an output

[SOURCE: EU AI Act 2024/1689 (Article 3(33))]

3.1.8
operator
provider (3.1.9), product manufacturer, *deployer* (3.1.4), authorized representative, importer or *distributor* (3.1.5)

[SOURCE: EU AI Act 2024/1689 (Article 3(8))]

3.1.9
provider
natural or legal person, public authority, agency or other body that develops an *AI system* (3.1.1) or a general-purpose *AI model* (3.2.1) or that has an *AI system* (3.1.1) or a general-purpose *AI model* (3.2.1) developed and places it on the market or puts the *AI system* (3.1.1) into service under its own name or trademark, whether for payment or free of charge

[SOURCE: EU AI Act 2024/1689 (Article 3(3)), modified - "a" at the beginning has been removed based on standard rules]

3.1.10
risk
combination of the *probability* (3.2.25) of an occurrence of harm and the severity of that harm

Note 1 to entry: there are slightly different definitions of *risk* (3.1.10) used by other European regulations, which include but not limited to NIS2, CRA etc.

[SOURCE: EU AI Act, Art 3.3(2), modified — removing article “the” in front of “combination” to align with ISO IEC CEN CLC directive part 2 rules for definition starting without article; Note 1 to entry added.]

3.1.11
testing data
data used for providing an independent evaluation of the *AI system* (3.1.1) in order to confirm the expected performance of that system before its placing on the market or putting into service

[SOURCE: EU AI Act 2024/1689 (Article 3(32))]

3.1.12
training data
data used for training an *AI system* (3.1.1) through fitting its learnable parameters

[SOURCE: EU AI Act 2024/1689 (Article 3(29))]

3.1.13
validation data
data used for providing an evaluation of the trained *AI system* (3.1.1) and for tuning its non-learnable parameters and its learning process in order, inter alia, to prevent underfitting or overfitting

[SOURCE: EU AI Act 2024/1689 (Article 3(30))]

3.2 Other related terms

3.2.1
AI model
computational model that generates outputs such as predictions, recommendations, classifications, or decisions based on *input data* (3.1.7)

Note 1 to entry: An *AI model* (3.2.1) can be trained using *machine learning* (3.2.26) techniques or other computational methods.

Note 2 to entry: An *AI model* (3.2.1) forms part of an *AI system* (3.1.1) together with other components such as data processing, *inference* (3.2.21) mechanisms and system interfaces.

3.2.2
adversarial example
input that includes a small perturbation compared to a typical or expected input, which results in a significantly different and unwanted behaviour of the *AI system* (3.1.1)

EXAMPLE Manipulating image recognition systems with artificial noise patterns to mislead classification or detection.

Note 1 to entry: Adversarial attacks rely on carefully crafting or selecting *adversarial example* (3.2.2).

3.2.3
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 [2], 3.2.]

3.2.4
availability
property of being accessible and usable on demand by an authorized entity

[SOURCE: ISO/IEC 27000:2018 [2], 3.7.]

3.2.5
circumstances
specific situational, short-term, internal and external condition of an *AI system* (3.1.1) in its *environment* (3.2.16), that enables or shapes *risk* (3.1.10) in this situation

3.2.6
confidentiality
property that information is not made available or disclosed to unauthorized individuals, entities, or processes

[SOURCE: ISO/IEC 27000:2018 [2], 3.10.]

3.2.7
confidentiality attack
cybersecurity *attack* (3.2.3) that compromises the *confidentiality* (3.2.6) of an *AI system* (3.1.1) for instance by extracting, inferring, or reconstructing sensitive information related to *training data* (3.1.12), model parameters, or proprietary system behaviour including proprietary system properties such as system prompts or embedded instructions

Note 1 to entry: confidentiality attacks include model inversion, membership *inference* (3.2.21), and model extraction.

Note 2 to entry: confidentiality attacks exploit observable system outputs, responses, or interfaces.

3.2.8
configuration
set of parameters, system prompts, execution settings, and integration settings that influence model behaviour

3.2.9
control
measure (3.2.27) intended to limit, contain, or manage the consequences of cybersecurity threats or attacks over time

Note 1 to entry: A *control* (3.2.9) can be technical or organisational and can operate continuously or conditionally.

Note 2 to entry: Controls support continued operation of the *AI system* (3.1.1) within acceptable risk levels by reducing propagation, recurrence, or systemic impact of attacks.

3.2.10
conventional cybersecurity
non-AI-specific *cybersecurity* (3.1.2)

3.2.11
cybersecurity incident for AI systems
cybersecurity incident
event or series of events that compromise, or have the potential to compromise, the *confidentiality* (3.2.6), *integrity* (3.2.24), or *availability* (3.2.4) of the *AI system* (3.1.1) or its assets

Note 1 to entry: In this document, the term "incident" refers to a *cybersecurity incident* (3.2.11) unless explicitly specified otherwise.

Note 2 to entry: A cybersecurity event is any observable occurrence related to the operation of an *AI system* (3.1.1).

Note 3 to entry: A cybersecurity event becomes a *cybersecurity incident* (3.2.11) when it results in, or could result in, adverse cybersecurity consequences.

[SOURCE: ISO/IEC 27000:2018 [2], 3.21, modified — the word "cybersecurity" has been added to clarify the scope of the term and cybersecurity incident was added as admitted term.]

3.2.12
data poisoning attack
causative attack
attack (3.2.3) where the attacker manipulates data used to train the model, affecting the AI system's behaviour including poisoning of data used during *inference* (3.2.21) or retrieval processes

Note 1 to entry: This *attack (3.2.3)* can also be referred to in ISO/IEC 19761 as "data manipulation" meaning "any processing of the data other than a movement of the data into or out of a functional *process*, or between a functional *process* and persistent storage" [ISO/IEC 19761, 2.6].

[SOURCE: ISO/IEC DIS 27090 [3]:—, 3.12, modified — Note 1 to entry added. Replace "data that the model uses to train to affect the AI system's behaviour" with "data used to train the model, affecting the AI system's behaviour".]

3.2.13
deployment component
component responsible for integrating and operating a trained *AI model* (3.2.1) within a production or operational *environment* (3.2.16)

3.2.14
detect
measure (3.2.27) intended to identify the occurrence or attempted occurrence of cybersecurity threats or attacks affecting the *AI system* (3.1.1)

Note 1 to entry: Detection is based on monitoring, analysis of system behaviour, data characteristics, or deviations from expected operation.

Note 2 to entry: Detection can occur during training, deployment, or operation of the *AI system* (3.1.1).

3.2.15
direct prompt injection
attack (3.2.3) where an attacker provides crafted input containing instructions directly to an *AI system* (3.1.1) in order to influence its behaviour contrary to its intended use

Note 1 to entry: *direct prompt injection* (3.2.15) differs from *indirect prompt injection* (3.2.20) in that the instructions are supplied explicitly as part of the input at the time of interaction with the *AI system* (3.1.1) .

Note 2 to entry: *direct prompt injection* (3.2.15) primarily affects AI systems that interpret user-provided inputs as instructions, including generative AI systems.

3.2.16
environment
sum of all structural and long-term factors (physical, digital, organizational) of an *AI system* (3.1.1) that influence the cybersecurity

Note 1 to entry: *environment* (3.2.16) is the asset landscape which needs protection as prevention.

3.2.17
model evasion
attack (3.2.3) where the attacker is able to alter correct behaviour of an *AI system* (3.1.1) on specific input or classes of inputs

3.2.18
exposure-restricted data
data that represent an unacceptable risk when exposed to unauthorized parties

3.2.19
fine-tuning data
training data (3.1.12) used to further adapt a previously trained *AI model* (3.2.1) to improve performance or adjust the model to a specific task or domain

Note 1 to entry: *fine-tuning data* (3.2.19) are a subset of *training data* (3.1.12) used after an initial model training phase.

Note 2 to entry: Fine-tuning typically adjusts existing model parameters rather than training a model from scratch.

3.2.20
indirect prompt injection
attack (3.2.3) where an attacker embeds instructions within data that are later processed as input by an *AI system* (3.1.1) in order to influence its behaviour contrary to its intended use

Note 1 to entry: *indirect prompt injection* (3.2.20) differs from *direct prompt injection* (3.2.15) in that the instructions are not provided explicitly by the user at the time of interaction, but are introduced through data sources such as documents, web content, or retrieved information.

Note 2 to entry: *indirect prompt injection* (3.2.20) primarily affects AI systems that process natural language or other structured or unstructured content as part of their inputs, including generative AI systems.

3.2.21
inference

reasoning by which conclusions are derived from known premises

Note 1 to entry: In AI, a premise is either a fact, a rule, a *model*, a feature or raw data.

Note 2 to entry: The term "*inference* (3.2.21)" refers both to the *process* and its result.

[SOURCE: ISO/IEC 22989:2022 [4], 3.1.17.]

3.2.22
inference component
component of an *AI system* (3.1.1) responsible for executing a trained *AI model* (3.2.1) to generate outputs based on *input data* (3.1.7)

Note 1 to entry: The *inference component* (3.2.22) processes *inference data* (3.2.23) and produces predictions, classifications, recommendations or other outputs depending on the *AI system* (3.1.1) function.

3.2.23
inference data
data provided to an *AI model* (3.2.1) during its operational phase as part of *input data* (3.1.7) for generating outputs

Note 1 to entry: This can include user inputs, system prompts, sensor data or contextual information.

3.2.24
integrity
property of accuracy and completeness

[SOURCE: ISO/IEC 27000:2018 [2], 3.36.]