centro criptológico nacional

ICT Security Guide
CCN-STIC 861

Applicability Criteria for ENS Security Measures against Offensive AI



Courtesy translation

November 2026

Courtesy translation

Edit:





Firmado digitalmente por CCN
Fecha: 2026.11.20 10:10:32 +01'00'

© National Cryptologic Centre (CCN), 2026

Publication date: November 2026

LIMITATION OF LIABILITY

This document is provided in accordance with the terms contained therein, expressly rejecting any type of implied warranty that may be found related. In no case may the National Cryptologic Center be held liable for direct, indirect, incidental or extraordinary damage arising from the use of the information and software indicated even when it is warned of such possibility.

LEGAL NOTICE

It is strictly forbidden, without the written authorization of the National Cryptologic Center, under the sanctions established in the laws, the partial or total reproduction of this document by any means or procedure, including reprography and computer treatment, and the distribution of copies of it through public rental or loan.

INDEX

1 INTRODUCTION 4
2 PURPOSE, SCOPE AND AUDIENCE 4
2.1 Purpose 4
2.2 Scope 5
2.3 Audience 5
3 CONCEPTS 5
4 GENERAL APPLICABILITY CRITERIA 7
4.1 Applicability criteria for ENS measures 7
4.1.1 [op.pl.1] – Risk analysis 8
4.1.2 [op.acc.3] – Segregation of functions and tasks 8
4.1.3 [op.acc.4] – Access rights management process 9
4.1.4 [op.acc.5] – Authentication mechanism (external users) 9
4.1.5 [op.acc.6] – Authentication mechanism (organization users) 9
4.1.6 [op.exp.1] – Asset inventory 10
4.1.7 [op.exp.2] – Security configuration 10
4.1.8 [op.exp.4] – Security maintenance and updates 10
4.1.9 [op.exp.6] – Protection against malicious code 11
4.1.10 [op.exp.7] – Incident management 11
4.1.11 [op.exp.8] – Recording of the activity 12
4.1.12 [op.ext.3] – Supply chain protection 12
4.1.13 [op.cont.1] – Impact analysis 12
4.1.14 [op.cont.2] – Continuity plan 13
4.1.15 [op.cont.3] – Periodic tests 13
4.1.16 [op.mon.1] – Intrusion detection 13
4.1.17 [op.mon.3] – Monitoring 14
4.1.18 [mp.per.3] – Awareness 14
4.1.19 [mp.per.4] – Training 15
4.1.20 [mp.com.4] – Separation of information flows in the network 15
4.1.21 [mp.sw.1] – IT applications development 15
4.1.22 [mp.sw.2] – Acceptance and commissioning 16
4.1.23 [mp.s.1] – Email protection 16
4.1.24 [mp.s.2] – Protection of web services and applications 16
4.1.25 [mp.s.4] – Protection against denial of service 17
4.2 Matrix of scenarios and related measures 17
5 PRIORITIZATION OF ACTIONS 19
6 DOCUMENTATION, MONITORING AND REVIEW 19

Courtesy translation

1 INTRODUCTION

The National Security Framework (*Esquema Nacional de Seguridad*, ENS), regulated by Royal Decree 311/2022, establishes the basic principles, minimum requirements and security measures intended to adequately protect the information processed and the services provided by the systems within its scope. Its application is based on risk management; prevention, detection, response and retention; lines of defense; continuous monitoring; and periodic reassessment.

Artificial intelligence (AI) used for offensive purposes changes the threat landscape by increasing the speed, scale, personalization, automation, precision and adaptability of well-known activities, such as reconnaissance, social engineering, vulnerability discovery and exploitation, the generation and mutation of malicious code, evasion of security controls, identity abuse and the chaining of actions. It may also facilitate the malicious exploitation of models, agents, plugins, interfaces, knowledge repositories and AI services integrated in the system itself or in its supply chain.

Protection should not depend on establishing whether a specific attack has used AI. Instead, the assessment should focus on whether the evolution of such capabilities affects the likelihood, impact, reaction times or effectiveness of safeguards. The response should primarily consist of adapting and verifying controls already established under the ENS, without automatically introducing new measures merely due to the presence of a particular technology.

This guide takes CCN-CERT BP/36, *Best Practices against the Offensive AI Model*, as a reference and complements its approach by establishing verifiable applicability criteria for the measures set out in Annex II of the ENS.

In this context, second additional provision of Royal Decree 311/2022 provides for the National Cryptologic Centre (CCN) to develop and disseminate information and communication technology security guidelines to facilitate compliance with its provisions. This guide falls within that technical guidance function and sets out criteria for the application of the measures in Annex II to risks arising from offensive artificial intelligence, without thereby modifying the catalogue of measures or the enforceability regime established in the ENS.

2 PURPOSE, SCOPE AND AUDIENCE

2.1 Purpose

This guide aims to establish **criteria for the application and adaptation of ENS security measures** so that information systems can address the evolving use of AI for offensive purposes in a consistent, proportionate, verifiable and risk-based manner.

The criteria set out in this guide **neither replace nor exempt the categorization of information systems** in accordance with the ENS or from carrying the corresponding risk analysis and risk management**.** The enforceability of security measures shall be determined by the system category, the security dimensions, the corresponding levels and, where appropriate, the risk analysis. The criteria set out in this guide make it possible to determine when the implementation of an applicable measure must specifically consider the risks and needs associated with the offensive AI threat.

Courtesy translation

2.2 Scope

The applicability criteria have a **specific and complementary scope**, limited to providing guidance on the applicability, adaptation or, where appropriate, reinforcement of certain security measures affected by risks arising from offensive artificial intelligence. They cover AI-enhanced attacks, even where the protected system does not incorporate artificial intelligence components, as well as the malicious exploitation of AI models, agents, plugins and services integrated into the system itself or its supply chain.

This document does not cover the **general obligations applicable to artificial intelligence systems**, including those relating to governance, inventory, intended purpose, quality, bias, performance, human oversight, overall security and the comprehensive lifecycle management of AI systems and models. These aspects will be addressed in a separate document. Criteria concerning AI agents and components within the system itself are limited to risks arising from their offensive exploitation (agent hijacking, prompt injection, misuse of delegated credentials, etc.).

2.3 Audience

This guide is intended for **entities and organizations within the scope of application of the ENS**, as well as for those involved in risk analysis and risk management and for the implementation, operation and supervision of security measures.

The ENS applies not only to public sector information systems, but also to **private sector** entities that, under a contractual relationship, provide services to public sector entities in the exercise of their competences, as set out in Article 2.3 of Royal Decree 311/2022. Furthermore, any organization may choose to apply the ENS voluntarily as a reference framework for improving the security of its information systems.

3 CONCEPTS

To ensure a consistent interpretation of the applicability criteria, the terms set out in the following table shall, for the purpose of this guide, have the meaning specified therein, without prejudice to the definitions established in the ENS and its implementing regulations.

Concept	Operational definition
Offensive AI	Use of AI systems, including generative models, language models and autonomous agents, to support, automate or execute phases of the attack cycle, such as target reconnaissance, phishing generation, malicious code development, vulnerability identification and exploitation, or operational decision-making.
AI agent	Component capable of interpreting objectives, planning or selecting actions, and invoking tools, interfaces or services with varying degrees of autonomy.
Non-human identity	Identity assigned to a service, workload, process, API, agent or software component.
Sensitive action	Action whose improper execution could significantly affect the confidentiality, integrity, availability, authenticity or

courtesy translation

Concept	Operational definition
	accountability of the system and which, therefore, requires segregation of duties, independent verification and logging.
Untrusted source	External or unverified content that may influence a model or agent, including documents, web pages, emails, messages, tool outputs and knowledge repositories.
Prompt injection	Direct or indirect manipulation of inputs or content to alter the intended behaviour of a model or agent, bypass restrictions or trigger unauthorized actions.
Excessive agency	Ability of an agent to access more information, tools or actions than necessary, or to act without limits and checks proportionate to the potential impact.
AI component	AI-based model, agent, plugin, connector, interface, knowledge repository or service that forms part of the system or of its supply chain.
Attack surface	Set of assets, services, interfaces, identities and components, including AI and third-party components, that an attacker may identify, abuse or exploit.
Attack cycle	Sequence of phases of an intrusion: reconnaissance, initial access, exploitation, privilege escalation, lateral movement, evasion, and actions on objectives or exfiltration.
Synthetic content	Text, audio, images or video generated or manipulated using AI to impersonate individuals or organisations, or to mimic legitimate communications, such as cloned voices or audiovisual impersonation (deepfakes).
Advanced impersonation	Impersonation of individuals, roles, suppliers or services through hyper-personalized communications or synthetic content, with the aim of obtaining credentials or information, or inducing improper actions.
Out-of-band verification	Verification of the requester's identity or authorization for a sensitive action, through a channel other than the one through which the request was received and not controlled by the requester.
Phishing-resistant multi-factor authentication	Multi-factor authentication whose factors cannot be intercepted, reused or relayed by a third party impersonating the legitimate service.
Delegated credential or privilege	Permission, credential or token through which a user or service enables an agent or tool to act on its behalf and within a defined scope.
Knowledge source	Repository of documents, databases or other content that a model or agent queries to generate responses or determine actions, such as those used in retrieval-augmented generation (RAG) systems.
Automated traffic management	Set of capabilities for identifying, limiting, challenging or blocking abusive automated traffic targeting services, web

translation

Concept	Operational definition
	applications and APIs, based on observed behaviour and without requiring evidence that AI is being used.
Detection use case	Rule, correlation or analysis designed to detect specific malicious behaviour based on system telemetry and logs.
Applicability criterion	Guideline indicating when and how to apply, adapt or strengthen an Annex II ENS measure or reinforcement against offensive AI, without modifying the catalogue of measures or its enforceability regime.
Compensatory measure	Alternative or interim safeguard that reduces risk until the intended measure can be implemented.
Residual risk	Risk remaining after safeguards have been applied, which must be identified, assessed and, where appropriate, accepted in accordance with the risk analysis.
Statement of Applicability	List of the Annex II ENS security measures selected for an information system, in accordance with its category and the risk analysis.

4 GENERAL APPLICABILITY CRITERIA

The applicability criteria provide a framework for determining, **in a proportionate and risk-based manner,** when it is necessary to apply, adapt or reinforce the ENS measures concerned, taking into account the characteristics of the system, the information and services being protected, their exposure to emerging threats and the evolution of technological capabilities. The aim is to facilitate a consistent and verifiable application of the ENS while preserving its **stability and technological neutrality**.

This guidance shall be applied **throughout the entire lifecycle of the affected system and services** from planning, design, development, acquisition and procurement to operation, monitoring, updating, continuity, replacement or decommissioning.

Rather than constituting an entirely new threat, **offensive artificial intelligence acts as a capability multiplier**: it increases the speed, scale, precision and degree of autonomy of existing offensive techniques, making these capabilities available to a wider range of threat actors.

The proposed applicability criteria make it possible to adapt Annex II of the ENS to a threat landscape in which artificial intelligence may increase the scale, speed, personalization and evasion capabilities of attacks, without altering the general approach to system categorization and the progressive application of security measures and reinforcements established by the ENS. They also make it possible to incorporate prevention, detection, verification and response capabilities against AI-enhanced threats, without replacing the risk-based approach and without the need to prove that the attacker has used artificial intelligence.

4.1 Applicability criteria for ENS measures

For each security measure, the **specific applicability criteria in relation to offensive AI** and their rationale are provided. The selection of security measures included in this guide is not exhaustive: measures not included shall remain applicable where appropriate in accordance with the ENS.

Courtesy translation

The criteria set out in this guide do not, by themselves, determine whether an ENS security measure is required. This shall be determined by the system category, the security dimensions, the corresponding levels and, where appropriate, the risk analysis.

The selection is not exhaustive: measures not included shall remain applicable where appropriate in accordance with the ENS. The presence of AI does not automatically raise the system category or, by itself, trigger all the applicability criteria.

4.1.1 [op.pl.1] – Risk analysis

Code	ENS measure
[op.pl.1]	**Risk analysis**

Applicability criterion
The risk analysis measure shall apply with the following particularities: (a) the identification of threats carried out in accordance with **[op.pl.1.2]** shall take into account those associated with the offensive use of artificial intelligence; (b) the identification of the main residual risks in accordance with **[op.pl.1.4]** shall take into account those associated with the offensive use of artificial intelligence.

Rationale
AI can reduce the time, knowledge and resources required to identify targets, discover vulnerabilities, tailor deception, generate malicious code and adapt attacks. This evolution affects the likelihood, speed, scale and simultaneity of attacks; the risk analysis shall therefore be reviewed whenever offensive capabilities, system exposure or the effectiveness of the implemented safeguards change.

4.1.2 [op.acc.3] – Segregation of functions and tasks

Code	ENS measure
[op.acc.3]	**Segregation of functions and tasks**

Applicability criterion
For sensitive actions, the application of **[op.acc.3.1] and [op.acc.3.2]** shall include the segregation of the capabilities to request, authorize, execute and review, including where artificial intelligence agents are involved. Approvals shall be assigned to independent identities that are not under the control of the same agent or the same credentials; where appropriate, human intervention shall be required, and authorizations and exceptions shall be logged.

Rationale
Offensive AI may accelerate identity compromise and the chaining of actions, increasing the impact of concentrating privileges in a single account, service or agent. Segregation prevents a compromised entity from independently completing critical operations on its own or evading oversight.

Courtesy translation

4.1.3 [op.acc.4] – Access rights management process

Code	ENS measure
[op.acc.4]	Access rights management process

Applicability criterion
The application of **[op.acc.4.1], [op.acc.4.2], [op.acc.4.3] and [op.acc.4.4]** shall require that explicit authorization, deny by default, minimum privilege, need to know and periodic review of permissions to also apply to agents, services and workloads, including technical accounts, non-human identities, delegated privileges and agents and tools permissions.

Rationale
Offensive AI may facilitate identity compromise, privilege escalation and the automated use of accounts. Limiting permissions reduces the scope of compromise of an identity or agent.

4.1.4 [op.acc.5] – Authentication mechanism (external users)

Code	ENS measure
[op.acc.5]	Authentication mechanism (external users)

Applicability criterion
Where required by the risk analysis, the authentication of external users shall prioritise phishing-resistant multi-factor authentication (MFA) mechanisms, complemented by limiting authentication attempts and detecting anomalous behaviour.

Rationale
Offensive AI facilitates highly personalized phishing, automated social engineering, credential guessing and large-scale access campaigns. Phishing-resistant authentication factors reduce the effectiveness of these attacks.

4.1.5 [op.acc.6] – Authentication mechanism (organization users)

Code	ENS measure
[op.acc.6]	Authentication mechanism (organization users)

Applicability criterion
Where required by the risk analysis, and particularly for privileged access, remote access or access to sensitive information, authentication mechanisms shall prioritise phishing-resistant multi-factor authentication, strengthen identity verification and credential recovery processes, and enable the detection of anomalous use of accounts and sessions. Where artificial intelligence agents or tools act on behalf of users within the organization, shared credentials and unrestricted delegations shall be avoided, the link to the identity authorizing the action shall be preserved, and and delegated access shall be revocable.

Rationale

courtesy translation

Offensive AI can facilitate staff impersonation, personalized phishing, automated access attempts and the misuse of credentials obtained through deception.

4.1.6 [op.exp.1] – Asset inventory

Code	ENS measure
[op.exp.1]	**Asset inventory**

Applicability criterion
Where the system incorporates or depends on artificial intelligence models, agents, plugins, knowledge sources, interfaces, tools or services with access or action capabilities, the inventory maintained in accordance with **[op.exp.1.1]** shall include them, together with their version, owner and dependencies. For third-party software components, the formal list required by reinforcement **[op.exp.1.r4.1]** (list of software components) shall complement that inventory, without replacing the identification of the data, action capabilities, owners and dependencies of artificial intelligence components.

Rationale
These components form part of the attack surface. Although the security measure already covers these assets, the inventory shall provide a greater level of detail where their autonomy or external dependencies affect the risk.

4.1.7 [op.exp.2] – Security configuration

Code	ENS measure
[op.exp.2]	**Security configuration**

Applicability criterion
Security configuration shall minimize exposed services, functions, ports, permissions and components in order to reduce the opportunities for automated or AI-assisted reconnaissance, exploitation and lateral movement. Where the system incorporates AI agents or components, the application of **[op.exp.2.2]** shall include a review of their configurations, permissions, connectors, interfaces and tools, restricting them to strictly necessary tools and actions.

Rationale
Minimum functionality reduces the attack surface available for automated reconnaissance, exploitation and abuse.

4.1.8 [op.exp.4] – Security maintenance and updates

Code	ENS measure
[op.exp.4]	**Security maintenance and updates**

Applicability criterion

Courtesy translation

The procedure established in accordance with **[op.exp.4.2]** shall, in addition to the provisions therein, take into account the exposure of the asset, the criticality of the service, the existence of known active exploitation, the possibility of automated or AI-assisted exploitation, and the alerts or advisories issued by the competent bodies. Where the assessment indicates that artificial intelligence may accelerate the discovery or exploitation of a vulnerability and increase the risk to the system, remediation deadlines shall be proportionately shortened, taking into account the exposure and the effectiveness of the available safeguards.

Rationale
The prioritization of updates shall incorporate offensive AI as a factor affecting both risk and urgency. Offensive AI may increase the likelihood of exploitation, facilitate large-scale campaigns against exposed components and reduce the time available to respond.

4.1.9 [op.exp.6] – Protection against malicious code

Code	ENS measure
[op.exp.6]	**Protection against malicious code**

Applicability criterion
To comply with **[op.exp.6.1]**, mechanisms for the prevention of and response to malicious code shall also cover malicious code generated, obfuscated, adapted or mutated using AI techniques for offensive purposes. Their effectiveness shall be periodically verified against rapidly evolving techniques by reviewing coverage, telemetry, behavioural analysis, and update and response times.

Rationale
Controls shall remain effective against fast-changing or polymorphic variants, automated generation, obfuscation and AI-assisted evasion techniques.

4.1.10 [op.exp.7] – Incident management

Code	ENS measure
[op.exp.7]	**Incident management**

Applicability criterion
To comply with **[op.exp.7.1]**, the comprehensive process for dealing with incidents that may have an impact on system security shall include incidents involving the use of artificial intelligence techniques for offensive purposes or which affect AI-based systems, agents, services or components. Response times, defensive automation, coordination mechanism and action thresholds shall be reviewed for automated, multi-stage and high-volume campaigns. Specific response procedures shall also be defined for advanced impersonation, including suspension of the affected actions, verification of authorizations, preservation of evidence, containment and coordination with third parties.

Rationale

Courtesy translation

Offensive AI can accelerate the attack lifecycle, particularly the chaining of actions, and increase the simultaneity and personalization of attacks, reducing the time available for classification, containment and recovery.

4.1.11 [op.exp.8] – Recording of the activity

Code	ENS measure
[op.exp.8]	**Recording of the activity**

Applicability criterion
To comply with **[op.exp.8.1]**, where the system incorporates software agents or AI-based components capable of performing actions, invoking tools, accessing information or interacting with other systems, activity logs shall enable the identification of the entity or agent performing the action, the tool or interface used, the result obtained and, where appropriate, the associated human approval. The logs shall also preserve the link to the user or service that activates the agent and to the affected resources. For sensitive actions subject to out-of-band verification, the logs shall record the verification performed, the channel used, the identity providing confirmation, the result and the associated authorization, without retaining unnecessary sensitive data.

Rationale
Traceability shall enable the reconstruction of actions performed by AI components operating autonomously or semi-autonomous actions.

4.1.12 [op.ext.3] – Supply chain protection

Code	ENS measure
[op.ext.3]	**Supply chain protection**

Applicability criterion
The impact analysis performed in accordance with **[op.ext.3.1]** shall take into account accidental or deliberate incidents arising from the use by suppliers of AI-based systems, models, services, agents, connectors or components. Suppliers of models, data, artificial intelligence services, libraries, APIs, plugins and agents shall be inventoried, their provenance and update status shall be verified, and the analysis shall be extended to relevant subcontractors and dependencies.

Rationale
AI components and their dependencies may expand the attack surface, introduce manipulated code or data and grant external parties the ability to access information or take action.

4.1.13 [op.cont.1] – Impact analysis

Code	ENS measure
[op.cont.1]	**Impact analysis**

Courtesy translation

Applicability criterion
Where indicated by the risk analysis, the impact analysis shall consider risk concentrations and common dependencies and shall incorporate scenarios involving rapid, simultaneous and multi-component disruption arising from offensive AI.

Rationale
It makes it possible to anticipate the effects of offensive AI on continuity, as it may increase the speed, scale and simultaneity of attacks and affect multiple identities, applications or dependencies.

4.1.14 [op.cont.2] – Continuity plan

Code	ENS measure
[op.cont.2]	Continuity plan

Applicability criterion
Where required by the risk analysis, the continuity plan shall provide for isolation, alternatives arrangements and verified restoration; it shall address simultaneous compromises, secure recovery, coordination with third parties and the integrity of restored components; and ensure that both manual and automated measures are tested under realistic conditions.

Rationale
Strengthens preparedness and recovery capabilities in the face of faster, larger-scale or simultaneous incidents, as automated attacks can shorten decision-making and recovery times.

4.1.15 [op.cont.3] – Periodic tests

Code	ENS measure
[op.cont.3]	Periodic tests

Applicability criterion
Where required by the risk analysis, periodic tests shall progressively incorporate automated, adaptive and simultaneous scenarios in order to verify that continuity and recovery capabilities remain effective against adversaries with greater speed and operational capacbilities.

Rationale
It enables periodic verification of the effectiveness of continuity and recovery capabilities.

4.1.16 [op.mon.1] – Intrusion detection

Code	ENS measure
[op.mon.1]	Intrusion detection

Applicability criterion
Where indicated by the risk analysis, intrusion detection based on known threat patterns or indicators shall be complemented by behavioural analysis and correlation, and telemetry and use cases shall be

Courtesy translation

extended to cover rapid changes in tactics, evasion, distributed attacks and multi-stage attack sequences.

Rationale
Detection must identify malicious behaviour and automated activity even when the use of AI by the attacker cannot be proven.

4.1.17 [op.mon.3] – Monitoring

Code	ENS measure
[op.mon.3]	Monitoring

Applicability criterion
Where required by the risk analysis, anomaly detection, correlation and alert validation capabilities shall be established to address automated, adaptive or AI-enhanced attack patterns. **Reinforcement [op.mon.3.r3.1]** shall be applied considering that the threat detection system must be capable of detecting automated, adaptive or AI-enhanced threats. Detection use cases shall be implemented to identify and correlate indications of advanced impersonation and communications manipulation, integrating the resulting alerts with the monitoring of sensitive actions.

Rationale
Offensive AI can accelerate reconnaissance, vary techniques and adapt attacks to observed defences. Reinforcement R3 is therefore directly relevant, as it requires monitoring capable of detecting advanced threats, APTs and significant anomalies in network traffic.

4.1.18 [mp.per.3] – Awareness

Code	ENS measure
[mp.per.3]	Awareness

Applicability criterion
Staff awareness shall include periodic reminders of the security regulations referred to in **[mp.per.3.1]** as well as the identification of hyper-personalized impersonation campaigns, synthetic multimedia content and other AI-enhanced deceptive techniques used for malicious or fraudulent purposes, such as highly personalized phishing, cloned voices, audiovisual impersonation, manipulated content and false urgent instructions. Verifiable procedures for pausing, out-of-band verification and escalation shall be included, and their application shall be assessed through campaigns, exercises or behavioural evidence, and not merely by the ability to recognize deception.

Rationale
AI makes it possible to create more convincing and scalable deception; the effectiveness of the measure therefore depends on staff applying verification procedures.

Courtesy translation

4.1.19 [mp.per.4] – Training

Code	ENS measure
[mp.per.4]	Training

Applicability criterion
Training provided in accordance with **[mp.per.4.1]** shall include, where appropriate based on the role or the risk analysis, content on AI-enhanced threats, synthetic social engineering, advanced impersonation, detection of and response to AI-assisted incidents, and the secure use of AI tools in security functions. Training shall be tailored by role (users, operations, cybersecurity, development, architecture, procurement and audit) and shall cover automated attacks, telemetry, investigation, agent security and the validation of AI-assisted components or code; Practical training and exercises shall be updated as techniques evolve.

Rationale
Technical teams need to be familiar with these techniques in order to detect them and respond effectively.

4.1.20 [mp.com.4] – Separation of information flows in the network

Code	ENS measure
[mp.com.4]	Separation of information flows in the network

Applicability criterion
Where required by the risk analysis, network segments and authorized information flows shall be reviewed; administration, development, test and production environments shall be segregated, and communications by artificial intelligence components and agents shall be restricted to the necessary destinations and services. Where the risk so justifies, more granular segmentation shall be applied, traffic between network zones shall be monitored, and tests shall be carried out to confirm that the rules prevent lateral movement and unauthorized access without compromising continuity.

Rationale
Offensive AI can accelerate internal reconnaissance, lateral movement and propagation following the compromise of a component. Segregation limits the scope of the incident and hinders access to critical assets from lower-trust zones.

4.1.21 [mp.sw.1] – IT applications development

Code	ENS measure
[mp.sw.1]	IT applications development

Applicability criterion

Courtesy translation

Application development shall address AI-generated or AI-assisted code, automatically introduced dependencies, prompt injection from untrusted sources, agent actions, and input and output validation. The same requirements for human review, testing and minimum privilege shall apply to both conventional software and software incorporating artificial intelligence components.

Rationale
AI can generate insecure code, facilitate the discovery and exploitation of vulnerabilities and increase the autonomy of components.

4.1.22 [mp.sw.2] – Acceptance and commissioning

Code	ENS measure
[mp.sw.2]	**Acceptance and commissioning**

Applicability criterion
Acceptance and commissioning shall provide evidence of the resilience of the application to the AI-enhanced attack scenarios considered in the risk analysis, as well as of potential compromise vectors introduced by AI-assisted code or components; such evidence shall be checked against the acceptance criteria and the accepted residual risk. Where appropriate, testing shall include static and dynamic analysis, code review, penetration testing and automated or adaptive scenarios.

Rationale
Security acceptance criteria are verified before entry into production, in appropriate environments.

4.1.23 [mp.s.1] – Email protection

Code	ENS measure
[mp.s.1]	**Email protection**

Applicability criterion
Where indicates by the risk analysis, email protection shall strengthen email authentication, content and behavioural analysis, URL and attachment protection, and verification procedures, so as to remain effective against synthetic content and adaptive campaigns.

Rationale
AI increases the quality and personalization of phishing, impersonation and business email compromise, and facilitates the generation of malicious messages, links and attachments.

4.1.24 [mp.s.2] – Protection of web services and applications

Code	ENS measure
[mp.s.2]	**Protection of web services and applications**

Courtesy translation

Applicability criterion
Where required by the risk analysis, web services and applications shall be protected by deploying automated traffic management solutions capable of distinguishing legitimate human interactions from automated attacks supported or enhanced by AI-based tools used for offensive purposes, including, among others, mass data scraping, systematic service probing and attempts at unauthorized account access. In such cases, input validation, rate limiting, API protection, behavioural analysis and anti-automation testing shall also be applied.

Rationale
AI can increase the speed, scale and adaptability of attacks against web services. Detection and containment focus on malicious behavior, without the need to prove the use of AI.

4.1.25 [mp.s.4] – Protection against denial of service

Code	ENS measure
[mp.s.4]	Protection against denial of service

Applicability criterion
The technologies deployed in accordance with **[mp.s.4.2]** shall mitigate anomalous or abusive increases in traffic, including those generated using automation tools or artificial intelligence techniques used for offensive purposes. Capacity, limits, redundancy, anti-DDoS protection and coordination with providers shall be reviewed.

Rationale
AI can facilitate the reconnaissance, coordination and variation of offensive traffic; it is therefore advisable to assess whether it increases the volume, distribution or evasion capability of such traffic.

4.2 Matrix of scenarios and related measures

The following matrix provides a proposed mapping traceability between risk scenarios and security measures. **It does not replace the risk analysis, nor does it constitute an exhaustive list.**

Security measure **[op.pl.1] Risk analysis** covers all scenarios applicable to the information system, which shall be considered when identifying threats and assessing and reviewing risks, including residual risk. For each scenario, the matrix identifies the security measures most directly related to its prevention, detection, response or recovery.

Security measures **[op.mon.3] Monitoring and [mp.per.4] Training are also cross-cutting. [op.mon.3] Monitoring** ensured that knowledge of threats, techniques and offensive AI capabilities relevant to the system remains up to date, while **[mp.per.4] Training**, prepares technical and security personnel to identify, prevent and respond to them. Their application and reinforcement shall be determined according to the system category, the risk analysis and the functions performed by personnel.

Scenario	Description	Related measures
Phishing and AI-enabled impersonation	Use of AI to generate highly convincing communications, content or interactions intended to impersonate identities and	[op.acc.3] Segregation of functions and tasks [op.acc.5] Authentication mechanism (external users)

Courtesy translation

Scenario	Description	Related measures
	obtain credentials or information, or to induce improper actions.	[op.acc.6] Authentication mechanism (organization users) [mp.per.3] Awareness [mp.per.4] Training [mp.s.1] Email protection
Automated vulnerability exploitation	Use of AI to identify, prioritize and exploit vulnerabilities in an automated manner, reducing the time between their discovery and malicious exploitation.	[op.exp.2] Security configuration [op.exp.4] Security maintenance and updates [op.mon.1] Intrusion detection [mp.s.2] Protection of web services and applications
AI-generated or AI-adapted malicious code	Use of AI to generate, modify, debug, translate, obfuscate or mutate malicious code, produce variants at greater speed or adapt artifacts to evade detection and response mechanisms.	[op.exp.6] Protection against malicious code [op.mon.1] Intrusion detection [op.exp.7] Incident management [mp.com.4] Separation of information flows on the network
Prompt injection against AI agents	Introduction of malicious instructions, either directly or through untrusted sources, intended to alter the intended behaviour of an agent or induce it to perform unauthorized actions.	[op.acc.4] Access rights management process [op.exp.2] Security configuration [op.exp.8] Recording of the activity [mp.sw.1] IT applications development [mp.sw.2] Acceptance and commissioning
Manipulation of AI outputs	Generation or modification of outputs intended to cause undesired effects when processed by other components, applications, APIs or tools without sufficient validation.	[op.exp.2] Security configuration [mp.sw.1] IT applications development [mp.sw.2] Acceptance and commissioning
RAG sources poisoning	Manipulation or injection of malicious content into the knowledge sources used by a RAG system in order to influence its responses or actions.	[op.exp.1] Asset inventory [op.ext.3] Supply chain protection [mp.sw.1] IT applications development [mp.sw.2] Acceptance and commissioning
Permission abuse or excessive agency	Misuse by, or through, an AI agent of assigned tools, credentials or permissions to perform unauthorized actions or actions beyond the scope necessary.	[op.acc.3] Segregation of functions and tasks [op.acc.4] Access rights management process [op.exp.8] Recording of the activity [mp.com.4] Separation of information flows on the network
AI-enabled information exfiltration	Unauthorized acquisition or transfer of information accessible to the model or agent, including data, credentials, secrets, context, internal instructions or	[op.acc.4] Access rights management process [op.exp.2] Security configuration [op.exp.8] Recording of the activity [mp.sw.1] IT applications development

Courtesy translation

Scenario	Description	Related measures
	information retrieved from other systems.	[mp.com.4] Separation of information flows on the network
Supply chain compromise	Introduction or exploitation of compromised components, models, data, libraries, services, APIs or updates to affect the system or facilitate unauthorized access, manipulation or actions.	[op.exp.1] Asset inventory [op.ext.3] Supply chain protection [mp.sw.2] Acceptance and commissioning
Abusive consumption and economical denial	Automated or induced use of models, agents, APIs or tools to exhaust capacity, quotas or budget through mass requests, recursive calls, loops or high-cost tasks.	[op.exp.2] Security configuration [op.mon.1] Intrusion detection [mp.s.4] Protection against denial of service
Simultaneous or multi-stage compromise	Automated or coordinated chaining of multiple attack techniques and stages, or the simultaneous compromise of multiple identities, services or dependencies, reducing the time available for detection, containment and recovery.	[op.exp.7] Incident management [op.cont.1] Impact analysis [op.cont.2] Continuity plan [op.cont.3] Periodic tests [mp.com.4] Separation of information flows on the network

5 PRIORITIZATION OF ACTIONS

Once applicability has been determined, the entity shall prioritize actions according to the risk they reduce rather than their number, in proportion to the system category and the results of the risk analysis, considering at least the following criteria:

- **Exposure of assets and services**, active exploitation and potential for automation.
- **Impact** on the security dimensions and criticality of the information and services affected.
- **Capability of agents** or components to perform actions, and the scope of their permissions.
- **Third-party dependencies**, concentration and difficulty of replacement.
- **Understanding of the time** available to detect, decide, contain and recover.
- **Effectiveness**, urgency, feasibility, cost and implementation timeframe.
- **Residual risk and the compensatory measures** required while actions are being implemented.

Priority shall be given to actions that reduce critical risks, actively exploited risks or risks capable of causing rapid and widespread effects. Cost or difficulty alone shall not justify leaving an unaccepted risk untreated.

6 DOCUMENTATION, MONITORING AND REVIEW

Decisions on applicability, adaptation, priority and acceptance shall be justified in the risk analysis and, where appropriate, reflected in the **Statement of Applicability and the risk treatment plans.** Responsibilities, deadlines, evidence, dependencies, exceptions, compensatory measures and residual risk shall be documented.

The effectiveness of the actions taken shall be verified through logs, tests and exercises. Decisions shall be reviewed in the event of relevant changes occur in threats, exposure, architecture, permissions, AI components, knowledge sources, suppliers, incidents, or verifications results.

Courtesy translation