Guía de Seguridad de las TIC
CCN-STIC 861

Criterios de Aplicabilidad de las medidas del ENS frente a la IA ofensiva



Noviembre 2026

Edita:

centro criptológico nacional

© Centro Criptológico Nacional, 2026

Firmado digitalmente por CCN
Fecha: 2026.11.16 12:27:49 +01'00'

Fecha de Edición: noviembre de 2026

LIMITACIÓN DE RESPONSABILIDAD

El presente documento se proporciona de acuerdo con los términos en él recogidos, rechazando expresamente cualquier tipo de garantía implícita que se pueda encontrar relacionada. En ningún caso, el Centro Criptológico Nacional puede ser considerado responsable del daño directo, indirecto, fortuito o extraordinario derivado de la utilización de la información y software que se indican incluso cuando se advierta de tal posibilidad.

AVISO LEGAL

Quedan rigurosamente prohibidas, sin la autorización escrita del Centro Criptológico Nacional, bajo las sanciones establecidas en las leyes, la reproducción parcial o total de este documento por cualquier medio o procedimiento, comprendidos la reprografía y el tratamiento informático, y la distribución de ejemplares del mismo mediante alquiler o préstamo públicos.

ÍNDICE

1 INTRODUCCIÓN 4
2 OBJETO, ALCANCE Y DESTINATARIOS 4
2.1 Objeto 4
2.2 Alcance 5
2.3 Destinatarios 5
3 CONCEPTOS 5
4 CRITERIOS GENERALES DE APLICABILIDAD 7
4.1 Criterios de aplicabilidad de las medidas del ENS 8
4.1.1 [op.pl.1] – Análisis de riesgos 8
4.1.2 [op.acc.3] – Segregación de funciones y tareas 8
4.1.3 [op.acc.4] – Proceso de gestión de derechos de acceso 9
4.1.4 [op.acc.5] – Mecanismos de autenticación de usuarios externos 9
4.1.5 [op.acc.6] – Mecanismos de autenticación (usuarios de la organización) 9
4.1.6 [op.exp.1] – Inventario de activos 10
4.1.7 [op.exp.2] – Configuración de seguridad 10
4.1.8 [op.exp.4] – Mantenimiento y actualizaciones de seguridad 11
4.1.9 [op.exp.6] – Protección frente a código dañino 11
4.1.10 [op.exp.7] – Gestión de incidentes 12
4.1.11 [op.exp.8] – Registro de la actividad 12
4.1.12 [op.ext.3] – Protección de la cadena de suministro 13
4.1.13 [op.cont.1] – Análisis de impacto 13
4.1.14 [op.cont.2] – Plan de continuidad 13
4.1.15 [op.cont.3] – Pruebas periódicas 14
4.1.16 [op.mon.1] – Detección de intrusión 14
4.1.17 [op.mon.3] – Vigilancia 14
4.1.18 [mp.per.3] – Concienciación 15
4.1.19 [mp.per.4] – Formación 15
4.1.20 [mp.com.4] – Separación de flujos de información en la red 16
4.1.21 [mp.sw.1] – Desarrollo de aplicaciones 16
4.1.22 [mp.sw.2] – Aceptación y puesta en servicio 17
4.1.23 [mp.s.1] – Protección del correo electrónico 17
4.1.24 [mp.s.2] – Protección de servicios y aplicaciones web 17
4.1.25 [mp.s.4] – Protección frente a denegación de servicio 18
4.2 Matriz de escenarios y medidas relacionadas 18
5 PRIORIZACIÓN DE LAS ACTUACIONES 20
6 DOCUMENTACIÓN, SEGUIMIENTO Y REVISIÓN 20

1 INTRODUCCIÓN

El Esquema Nacional de Seguridad (ENS), regulado por el Real Decreto 311/2022, establece los principios básicos, requisitos mínimos y medidas de seguridad destinados a proteger adecuadamente la información tratada y los servicios prestados por los sistemas incluidos en su ámbito de aplicación. Su aplicación se basa en la gestión de riesgos, la prevención, detección, respuesta y conservación, la existencia de líneas de defensa, la vigilancia continua y la reevaluación periódica.

La inteligencia artificial (IA) utilizada con fines ofensivos modifica el contexto de amenaza al incrementar la velocidad, escala, personalización, automatización, precisión y capacidad de adaptación de actividades ya conocidas, como el reconocimiento, la ingeniería social, el descubrimiento y explotación de vulnerabilidades, la generación y mutación de código dañino, la evasión de controles, el abuso de identidades y el encadenamiento de acciones. También puede facilitar el aprovechamiento malicioso de modelos, agentes, complementos, interfaces, repositorios de conocimiento y servicios de IA incorporados al propio sistema o a su cadena de suministro.

La protección no debe depender de acreditar si un ataque concreto ha empleado IA. Debe valorarse si la evolución de esas capacidades altera la probabilidad, el impacto, los tiempos de reacción o la eficacia de las salvaguardas. La respuesta consiste, principalmente, en adaptar y verificar controles ya previstos por el ENS, sin introducir automáticamente nuevas medidas por la mera presencia de una tecnología.

La guía toma como referencia la CCN-CERT BP/36, Buenas prácticas frente al modelo de IA ofensiva, y complementa su enfoque mediante criterios de aplicabilidad verificables sobre las medidas del anexo II del ENS.

En este marco, la disposición adicional segunda del Real Decreto 311/2022 prevé que el Centro Criptológico Nacional (CCN) elabore y difunda guías de seguridad de las tecnologías de la información y las comunicaciones para facilitar el cumplimiento de sus disposiciones. La presente guía se encuadra en esa función de orientación técnica y concreta criterios para la aplicación de las medidas del anexo II ante los riesgos derivados de la inteligencia artificial ofensiva, sin que ello suponga modificar por sí misma el catálogo de medidas ni el régimen de exigibilidad establecido en el ENS.

2 OBJETO, ALCANCE Y DESTINATARIOS

2.1 Objeto

Esta guía tiene por objeto establecer **criterios para la aplicación y adaptación de las medidas del ENS** de manera que los sistemas de información puedan hacer frente a la evolución de la IA con fines ofensivos de forma homogénea, proporcionada y verificable basada en el riesgo.

Los criterios establecidos en esta guía **no sustituyen ni eximen de la categorización de los sistemas** de información conforme al ENS ni de la realización del correspondiente análisis y gestión de riesgos. La exigibilidad de las medidas vendrá determinada por la categoría del sistema, las dimensiones de seguridad, los niveles correspondientes y, cuando proceda, el análisis de riesgos. Los criterios establecidos en esta guía permiten determinar cuándo la implantación de una medida aplicable debe considerar específicamente los riesgos y necesidades asociados a la amenaza de la IA ofensiva.

2.2 Alcance

Los criterios de aplicabilidad tienen un **alcance específico y complementario**, limitado a orientar la aplicabilidad, adaptación o, cuando proceda, refuerzo de determinadas medidas de seguridad afectadas por los riesgos derivados de la inteligencia artificial ofensiva. Comprenden los ataques potenciados por inteligencia artificial, aunque el sistema protegido no incorpore componentes de inteligencia artificial, y el aprovechamiento malicioso de modelos, agentes, complementos y servicios de inteligencia artificial incorporados al propio sistema o a su cadena de suministro.

Quedan fuera del alcance de este documento las **obligaciones generales propias de los sistemas de inteligencia artificial**, incluidas las relativas a la gobernanza, inventario, finalidad, calidad, sesgo, rendimiento, supervisión humana, seguridad integral y gestión completa del ciclo de vida de sistemas y modelos de inteligencia artificial, que serán objeto de un documento independiente. Los criterios sobre agentes y componentes de IA del propio sistema se limitan a los riesgos de su aprovechamiento ofensivo (secuestro de agentes, inyección de instrucciones, abuso de credenciales delegadas…).

2.3 Destinatarios

La guía está dirigida a las **entidades y organizaciones comprendidas en el ámbito de aplicación del ENS,** así como a los responsables que intervienen en el análisis y gestión de riesgos y en la implantación, operación y supervisión de las medidas de seguridad.

El ENS no solo es de aplicación a los sistemas de información del sector público, sino también a aquellos del **sector privado** que, en virtud de una relación contractual, presten servicios competenciales a las entidades del sector público, tal y como establece el artículo 2.3 del Real Decreto 311/2022. Asimismo, cualquier organización puede optar por aplicar voluntariamente el ENS como marco de referencia para mejorar la seguridad de sus sistemas.

3 CONCEPTOS

Para facilitar una interpretación homogénea de los criterios de aplicabilidad, los términos recogidos en la siguiente tabla se entenderán, a efectos de esta guía, con el significado que se indica, sin perjuicio de las definiciones establecidas en el ENS y en su normativa de desarrollo.

Concepto	Definición operativa
IA ofensiva	Uso de sistemas de IA, incluidos modelos generativos, modelos de lenguaje y agentes autónomos, para apoyar, automatizar o ejecutar fases del ciclo de ataque, como el reconocimiento de objetivos, la generación de phishing, el desarrollo de código dañino, la identificación y explotación de vulnerabilidades o la toma de decisiones operativas
Agente de IA	Componente capaz de interpretar objetivos, planificar o seleccionar acciones e invocar herramientas, interfaces o servicios con distintos grados de autonomía.
Identidad no humana	Identidad asignada a un servicio, carga de trabajo, proceso, API, agente o componente software.
Actuación sensible	Actuación cuya ejecución indebida pueda afectar de forma significativa a la confidencialidad, integridad, disponibilidad,

Concepto	Definición operativa
	autenticidad o trazabilidad del sistema, y que, por ello, requiere segregación de funciones, verificación independiente y registro
Fuente no confiable	Contenido externo o no verificado que puede influir en un modelo o agente, incluidos documentos, páginas web, correos, mensajes, resultados de herramientas y repositorios de conocimiento.
Inyección de instrucciones	Manipulación directa o indirecta de entradas o contenidos para alterar el comportamiento previsto de un modelo o agente, eludir restricciones o provocar acciones no autorizadas.
Agencia excesiva	Capacidad de un agente para acceder a más información, herramientas o acciones de las necesarias, o para actuar sin límites y verificaciones proporcionales al impacto.
Componente de IA	Modelo, agente, complemento, conector, interfaz, repositorio de conocimiento o servicio basado en IA que forma parte del sistema o de su cadena de suministro.
Superficie de ataque	Conjunto de activos, servicios, interfaces, identidades y componentes, incluidos los de IA y de terceros, que un atacante puede reconocer, abusar o explotar.
Ciclo de ataque	Secuencia de fases de una intrusión: reconocimiento, acceso inicial, explotación, escalada de privilegios, movimiento lateral, evasión y actuación sobre los objetivos o exfiltración.
Contenido sintético	Texto, audio, imagen o vídeo generado o manipulado mediante IA que imita a personas, organizaciones o comunicaciones legítimas, como las voces clonadas o la suplantación audiovisual.
Suplantación avanzada	Suplantación de personas, cargos, proveedores o servicios mediante comunicaciones hiperpersonalizadas o contenido sintético, con el fin de obtener credenciales, información o la ejecución de acciones indebidas.
Verificación por canal independiente	Comprobación de la identidad del solicitante o de la autorización de una actuación sensible a través de un canal distinto del que recibió la solicitud y no controlado por quien la formula.
Autenticación multifactor resistente al phishing	Autenticación multifactor cuyos factores no pueden ser interceptados, reutilizados ni retransmitidos por un tercero que suplante el servicio legítimo.
Credencial o privilegio delegado	Permiso, credencial o token con el que un usuario o servicio habilita a un agente o herramienta para actuar en su nombre y dentro de un alcance determinado.
Fuente de conocimiento	Repositorio de documentos, bases de datos u otros contenidos que un modelo o agente consulta para generar respuestas o decidir acciones, como los empleados en sistemas de generación aumentada por recuperación (RAG).
Gestión de tráfico automatizado	Conjunto de capacidades para identificar, limitar, desafiar o bloquear tráfico automatizado abusivo hacia servicios,

Concepto	Definición operativa
	aplicaciones web y API, basadas en el comportamiento y sin necesidad de acreditar el uso de IA.
Caso de uso de detección	Regla, correlación o análisis diseñado para detectar un comportamiento malicioso concreto a partir de la telemetría y los registros del sistema.
Criterio de aplicabilidad	Pauta que orienta cuándo y cómo aplicar, adaptar o reforzar una medida o refuerzo del anexo II del ENS frente a la IA ofensiva, sin modificar el catálogo de medidas ni su régimen de exigibilidad.
Medida compensatoria	Salvaguarda alternativa o provisional que reduce el riesgo mientras no pueda implantarse la actuación prevista.
Riesgo residual	Riesgo que permanece tras aplicar las salvaguardas y que debe ser identificado, valorado y, en su caso, aceptado conforme al análisis de riesgos.
Declaración de Aplicabilidad	Relación de las medidas de seguridad del anexo II del ENS seleccionadas para un sistema, de acuerdo con su categoría y con el análisis de riesgos.

4 CRITERIOS GENERALES DE APLICABILIDAD

Los criterios de aplicabilidad proporcionan un marco para determinar, de **forma proporcionada y basada en el riesgo,** cuándo resulta necesario aplicar o reforzar las medidas del ENS afectadas, teniendo en cuenta las características del sistema, la información y los servicios protegidos, su exposición a las nuevas amenazas y la evolución de las capacidades tecnológicas. Con ello se pretende facilitar una aplicación homogénea y verificable del ENS, preservando al mismo tiempo su **estabilidad y neutralidad tecnológica**.

Las orientaciones deberán aplicarse **durante todo el ciclo de vida del sistema** y de los servicios afectados, desde la planificación, adquisición y contratación hasta su operación, supervisión, actualización, continuidad, sustitución o retirada.

La **inteligencia artificial ofensiva** más que una amenaza completamente nueva, **representa un multiplicador de capacidades**: aumenta la velocidad, la escala, la precisión y el grado de autonomía de técnicas ofensivas ya conocidas, facilitando estas capacidades a un mayor número de actores.

Los criterios de aplicabilidad propuestos permiten adaptar el Anexo II del ENS a un contexto de amenazas en el que la inteligencia artificial puede incrementar la escala, velocidad, personalización y capacidad de evasión de los ataques, sin alterar la lógica general de categorización y aplicación progresiva de medidas y refuerzos prevista en el ENS. Asimismo, permiten incorporar capacidades de prevención, detección, verificación y respuesta frente a amenazas potenciadas por inteligencia artificial, sin sustituir el enfoque basado en el riesgo y sin que sea necesario acreditar el empleo de inteligencia artificial por el atacante.

4.1 Criterios de aplicabilidad de las medidas del ENS

Para cada medida se indican los **criterios de aplicabilidad específicos frente a la IA ofensiva** y su justificación. La selección de medidas recogidas en esta guía no tiene carácter exhaustivo. Las medidas no incluidas seguirán siendo aplicables cuando corresponda conforme al ENS.

Los criterios establecidos en esta guía no determinan por sí mismos la exigibilidad de una medida del ENS. Esta vendrá determinada por la categoría del sistema, las dimensiones de seguridad, los niveles correspondientes y, cuando proceda, el análisis de riesgos.

La selección no es exhaustiva: las medidas no incluidas seguirán siendo aplicables cuando corresponda conforme al ENS. La presencia de IA no eleva automáticamente la categoría del sistema ni activa por sí solo todos los criterios de aplicabilidad.

4.1.1 [op.pl.1] – Análisis de riesgos

Código	Medida del ENS
[op.pl.1]	**Análisis de riesgos**

Criterio de aplicabilidad
La medida de análisis de riesgos será de aplicación con las siguientes particularidades: (a) la identificación de amenazas realizada conforme a **[op.pl.1.2]** tendrá en cuenta las asociadas al uso ofensivo de la inteligencia artificial; (b) la identificación de los principales riesgos residuales conforme a **[op.pl.1.4]** tendrá en cuenta los asociados al uso ofensivo de la inteligencia artificial.

Justificación
La IA puede reducir el tiempo, conocimiento y recursos necesarios para reconocer objetivos, descubrir vulnerabilidades, personalizar engaños, generar código dañino y adaptar ataques. Esta evolución modifica la probabilidad, velocidad, escala y simultaneidad de los ataques, por lo que el análisis ha de revisarse cuando cambien las capacidades ofensivas, la exposición del sistema o la eficacia de las salvaguardas implantadas.

4.1.2 [op.acc.3] – Segregación de funciones y tareas

Código	Medida del ENS
[op.acc.3]	**Segregación de funciones y tareas**

Criterio de aplicabilidad
En las actuaciones sensibles, la aplicación de **[op.acc.3.1] y [op.acc.3.2]** incluirá la separación de las capacidades de solicitar, autorizar, ejecutar y revisar, también cuando intervengan agentes de inteligencia artificial. Las aprobaciones corresponderán a identidades independientes que no queden bajo el control del mismo agente o de las mismas credenciales y, cuando proceda, se exigirá intervención humana y se registrarán las autorizaciones y excepciones.

Justificación
La IA ofensiva puede acelerar el compromiso de identidades y el encadenamiento de acciones, aumentando el impacto de concentrar privilegios en una misma cuenta, servicio o agente. La segregación evita que una entidad comprometida complete por sí sola operaciones críticas o eluda su supervisión.

4.1.3 [op.acc.4] – Proceso de gestión de derechos de acceso

Código	Medida del ENS
[op.acc.4]	Proceso de gestión de derechos de acceso

Criterio de aplicabilidad
La aplicación de **[op.acc.4.1], [op.acc.4.2], [op.acc.4.3] y [op.acc.4.4]** requerirá que la autorización expresa, la denegación por defecto, el mínimo privilegio, la necesidad de conocer y la revisión periódica de permisos se apliquen también a agentes, servicios y cargas de trabajo, incluyendo las cuentas técnicas, las identidades no humanas, los privilegios delegados y los permisos de agentes y herramientas.

Justificación
La IA ofensiva puede facilitar el compromiso de identidades, el escalado de privilegios y el uso automatizado de cuentas. Limitar los permisos reduce el alcance de una identidad o agente comprometido.

4.1.4 [op.acc.5] – Mecanismos de autenticación de usuarios externos

Código	Medida del ENS
[op.acc.5]	Mecanismos de autenticación de usuarios externos

Criterio de aplicabilidad
Cuando así resulte del análisis de riesgos, la autenticación de usuarios externos priorizará mecanismos multifactor (MFA) resistentes al phishing y se complementará con la limitación de intentos y la detección de comportamientos anómalos.

Justificación
La IA ofensiva facilita el phishing altamente personalizado, la ingeniería social automatizada, la adivinación de credenciales y las campañas masivas de acceso. Los factores resistentes al phishing reducen su eficacia.

4.1.5 [op.acc.6] – Mecanismos de autenticación (usuarios de la organización)

Código	Medida del ENS
[op.acc.6]	Mecanismos de autenticación (usuarios de la organización)

Criterio de aplicabilidad
Cuando así resulte del análisis de riesgos, y especialmente en los accesos con derechos de administración, remotos o a información sensible, los mecanismos de autenticación priorizarán la autenticación multifactor resistente al phishing, reforzarán la verificación de identidad y la recuperación de credenciales y permitirán detectar el uso anómalo de cuentas y sesiones. Cuando agentes o herramientas de inteligencia artificial actúen por cuenta de usuarios de la organización, se evitarán las credenciales compartidas y las delegaciones ilimitadas, se conservará la vinculación con la identidad que autoriza la actuación y se permitirá la revocación de los accesos delegados.

Justificación
La IA ofensiva puede facilitar la suplantación del personal, el phishing personalizado, la automatización de intentos de acceso y el abuso de credenciales obtenidas mediante engaño.

4.1.6 [op.exp.1] – Inventario de activos

Código	Medida del ENS
[op.exp.1]	**Inventario de activos**

Criterio de aplicabilidad
Cuando el sistema incorpore o dependa de modelos, agentes, complementos, fuentes de conocimiento, interfaces, herramientas o servicios de inteligencia artificial con capacidad de acceso o actuación, el inventario mantenido conforme a **[op.exp.1.1]** los incluirá, con su versión, responsable y dependencias. Para los componentes software de terceros, la relación formal exigida por el refuerzo **[op.exp.1.r4.1]** Lista de componentes software complementará dicho inventario, sin sustituir la identificación de los datos, capacidades de actuación, responsables y dependencias de los componentes de inteligencia artificial.

Justificación
Estos componentes forman parte de la superficie de ataque. Aunque la medida ya comprende estos activos, el detalle del inventario aumenta cuando su autonomía o su dependencia externa modifican el riesgo.

4.1.7 [op.exp.2] – Configuración de seguridad

Código	Medida del ENS
[op.exp.2]	**Configuración de seguridad**

Criterio de aplicabilidad
La configuración de seguridad minimizará los servicios, funciones, puertos, permisos y componentes expuestos para reducir las posibilidades de reconocimiento, explotación y movimiento lateral automatizados o asistidos por IA.

Cuando el sistema incorpore agentes o componentes de IA, la aplicación de **[op.exp.2.2]** incluirá la revisión de sus configuraciones, permisos, conectores, interfaces y herramientas, limitándolos a las herramientas y acciones estrictamente necesarias.

Justificación
La mínima funcionalidad reduce la superficie disponible para el reconocimiento, la explotación y el abuso automatizados.

4.1.8 [op.exp.4] – Mantenimiento y actualizaciones de seguridad

Código	Medida del ENS
[op.exp.4]	**Mantenimiento y actualizaciones de seguridad**

Criterio de aplicabilidad
El procedimiento establecido conforme a **[op.exp.4.2]** tendrá en cuenta, adicionalmente a lo indicado, la exposición del activo, la criticidad del servicio, la existencia de explotación activa conocida, la posibilidad de explotación automatizada o asistida mediante inteligencia artificial y las alertas o avisos emitidos por los organismos competentes. Cuando la evaluación indique que la inteligencia artificial puede acelerar el descubrimiento o la explotación de una vulnerabilidad e incrementar el riesgo del sistema, los plazos de corrección se acortarán de forma proporcionada, atendiendo a la exposición y a la eficacia de las salvaguardas disponibles.

Justificación
La priorización de actualizaciones deberá incorporar la IA ofensiva como factor de variación del riesgo y de urgencia. La IA ofensiva puede aumentar la probabilidad de explotación, facilitar campañas masivas contra componentes expuestos y reducir el tiempo disponible para reaccionar.

4.1.9 [op.exp.6] – Protección frente a código dañino

Código	Medida del ENS
[op.exp.6]	**Protección frente a código dañino**

Criterio de aplicabilidad
Para dar cumplimiento a **[op.exp.6.1]** los mecanismos de prevención y reacción frente a código dañino incluirán aquel generado, ofuscado, adaptado o mutado mediante técnicas de inteligencia artificial empleadas con fines ofensivos. Se verificará periódicamente su eficacia ante cambios rápidos de técnicas, revisando la cobertura, la telemetría, el análisis de comportamiento y los tiempos de actualización y respuesta.

Justificación

Los controles deberán mantener su eficacia frente a variantes rápidas o polimórficas, generación automatizada, ofuscación y técnicas de evasión asistidas mediante IA.

4.1.10 [op.exp.7] – Gestión de incidentes

Código	Medida del ENS
[op.exp.7]	**Gestión de incidentes**

Criterio de aplicabilidad
Para dar cumplimiento a **[op.exp.7.1]** el proceso integral para hacer frente a los incidentes que puedan tener un impacto en la seguridad del sistema incluirá aquellos en los que se empleen técnicas de inteligencia artificial con fines ofensivos o que afecten a sistemas, agentes, servicios o componentes basados en inteligencia artificial. Se revisarán los tiempos de respuesta, la automatización defensiva, la coordinación y los umbrales de actuación frente a campañas automatizadas, multietapa y de gran volumen, y se definirán procedimientos específicos de respuesta ante suplantación avanzada: suspensión de las actuaciones afectadas, comprobación de autorizaciones, preservación de evidencias, contención y coordinación con terceros.

Justificación
La IA ofensiva puede acelerar el ciclo de ataque, en particular el encadenamiento de acciones, y aumentar la simultaneidad y personalización de los ataques, reduciendo el tiempo disponible para clasificar, contener y recuperar.

4.1.11 [op.exp.8] – Registro de la actividad

Código	Medida del ENS
[op.exp.8]	**Registro de la actividad**

Criterio de aplicabilidad
Para dar cumplimiento a **[op.exp.8.1]** se tendrá en cuenta que, cuando el sistema incorpore agentes software o componentes basados en inteligencia artificial con capacidad de ejecutar acciones, invocar herramientas, acceder a información o interactuar con otros sistemas, el registro de actividad deberá permitir identificar la entidad o agente que realiza la acción, la herramienta o interfaz utilizada, el resultado obtenido y, cuando proceda, la aprobación humana asociada. El registro conservará también la vinculación con el usuario o servicio que activa al agente y con los recursos afectados. En las actuaciones sensibles sujetas a verificación por canal independiente, el registro incluirá la verificación realizada, el canal utilizado, la identidad que confirma, el resultado y la autorización asociada, sin conservar datos sensibles innecesarios.

Justificación
La trazabilidad debe permitir reconstruir las actuaciones de los componentes de IA que ejecuten acciones autónomas o semiautónomas.

4.1.12 [op.ext.3] – Protección de la cadena de suministro

Código	Medida del ENS
[op.ext.3]	**Protección de la cadena de suministro**

Criterio de aplicabilidad
El análisis de impacto realizado conforme a **[op.ext.3.1]** tendrá en cuenta los incidentes accidentales o deliberados derivados del uso de sistemas, modelos, servicios, agentes, conectores o componentes basados en inteligencia artificial utilizados por proveedores. Se inventariarán los proveedores de modelos, datos, servicios de inteligencia artificial, bibliotecas, API, complementos y agentes, se verificará su procedencia y actualización y el análisis se extenderá a subcontratistas y dependencias relevantes.

Justificación
Los componentes de IA y sus dependencias pueden ampliar la superficie de ataque, introducir código o datos manipulados y otorgar capacidades externas de acceso o actuación.

4.1.13 [op.cont.1] – Análisis de impacto

Código	Medida del ENS
[op.cont.1]	**Análisis de impacto**

Criterio de aplicabilidad
Cuando así resulte del análisis de riesgos, el análisis de impacto considerará las concentraciones de riesgo y las dependencias comunes e incorporará escenarios de interrupción rápida, simultánea y multicomponente derivados de la IA ofensiva.

Justificación
Permite anticipar los efectos de la IA ofensiva sobre la continuidad, ya que puede aumentar la velocidad, escala y simultaneidad de los ataques y afectar a varias identidades, aplicaciones o dependencias.

4.1.14 [op.cont.2] – Plan de continuidad

Código	Medida del ENS
[op.cont.2]	**Plan de continuidad**

Criterio de aplicabilidad
Cuando así resulte del análisis de riesgos, el plan de continuidad preverá aislamiento, alternativas y restauración verificada, contemplará compromisos simultáneos, recuperación segura, coordinación con terceros e integridad de los componentes restaurados, y sus medidas manuales y automáticas se probarán en condiciones realistas.

Justificación
Refuerza la preparación y la capacidad de recuperación ante incidentes de mayor velocidad, escala o simultaneidad, pues los ataques automatizados pueden reducir los tiempos de decisión y recuperación.

4.1.15 [op.cont.3] – Pruebas periódicas

Código	Medida del ENS
[op.cont.3]	Pruebas periódicas

Criterio de aplicabilidad
Cuando así resulte del análisis de riesgos, las pruebas periódicas incorporarán progresivamente escenarios automatizados, adaptativos y simultáneos, para comprobar que las capacidades de continuidad y recuperación siguen siendo eficaces frente a adversarios con mayor velocidad y capacidad operativa.

Justificación
Permite verificar de forma periódica la eficacia de las capacidades de continuidad y recuperación.

4.1.16 [op.mon.1] – Detección de intrusión

Código	Medida del ENS
[op.mon.1]	Detección de intrusión

Criterio de aplicabilidad
Cuando así resulte del análisis de riesgos, la detección de intrusión basada en patrones o indicadores de amenazas conocidas se complementará con análisis de comportamiento y correlación y ampliará la telemetría y los casos de uso a cambios rápidos de táctica, evasión, ataques distribuidos y secuencias multietapa.

Justificación
La detección debe identificar comportamientos maliciosos y actividad automatizada aunque no pueda acreditarse el uso de IA por el atacante.

4.1.17 [op.mon.3] – Vigilancia

Código	Medida del ENS
[op.mon.3]	Vigilancia

Criterio de aplicabilidad

Cuando así resulte del análisis de riesgos, se establecerán capacidades de detección de anomalías, correlación y validación de alertas frente a patrones de ataque automatizados, adaptativos o potenciados mediante inteligencia artificial. El **refuerzo [op.mon.3.r3.1]** se aplicará teniendo en cuenta que el sistema de detección de amenazas debe ser capaz de detectar amenazas automatizadas, adaptativas o potenciadas mediante inteligencia artificial. Se incorporarán casos de uso para detectar y correlacionar indicios de suplantación avanzada y manipulación de comunicaciones, integrando las alertas con la supervisión de actuaciones sensibles.

Justificación
La IA ofensiva puede acelerar el reconocimiento, variar técnicas y adaptar los ataques a las defensas observadas. Por ello el refuerzo R3 resulta directamente pertinente, ya que la vigilancia detecta amenazas avanzadas, APT y anomalías significativas en el tráfico.

4.1.18 [mp.per.3] – Concienciación

Código	Medida del ENS
[mp.per.3]	**Concienciación**

Criterio de aplicabilidad
La concienciación del personal incluirá el recordatorio periódico de la normativa de seguridad indicada en **[mp.per.3.1]** y la identificación de campañas de suplantación de identidad hiperpersonalizadas, de contenidos multimedia sintéticos y de otros engaños potenciados por inteligencia artificial empleados con fines maliciosos o fraudulentos, como el phishing altamente personalizado, las voces clonadas, la suplantación audiovisual, el contenido manipulado y las falsas instrucciones urgentes. Se incluirán procedimientos verificables de pausa, verificación por canal independiente y escalado, y se evaluará su aplicación mediante campañas, ejercicios o evidencias de comportamiento, y no solo el reconocimiento del engaño.

Justificación
La IA permite crear engaños más convincentes y escalables, por lo que la eficacia de la medida depende de que el personal aplique procedimientos de comprobación.

4.1.19 [mp.per.4] – Formación

Código	Medida del ENS
[mp.per.4]	**Formación**

Criterio de aplicabilidad
La formación impartida conforme a **[mp.per.4.1]** incluirá, cuando proceda por razón del puesto o del análisis de riesgos, contenidos relativos a amenazas potenciadas por inteligencia artificial, ingeniería social sintética, suplantación avanzada, detección y respuesta ante incidentes asistidos por IA y uso seguro de herramientas de IA en funciones de seguridad. La formación se adaptará por rol (usuarios, operación, ciberseguridad, desarrollo, arquitectura, contratación y auditoría), incluirá ataques automatizados, telemetría, investigación, seguridad de

agentes y validación de componentes o código asistido por IA, y su componente práctico y sus simulacros se actualizarán conforme evolucionen las técnicas.

Justificación
Los equipos técnicos necesitan conocer estas técnicas para detectarlas y responder con eficacia.

4.1.20 [mp.com.4] – Separación de flujos de información en la red

Código	Medida del ENS
[mp.com.4]	**Separación de flujos de información en la red**

Criterio de aplicabilidad
Cuando así resulte del análisis de riesgos, se revisarán los segmentos y los flujos autorizados, se separarán los entornos de administración, desarrollo, pruebas y producción y se restringirán las comunicaciones de los componentes y agentes de inteligencia artificial a los destinos y servicios necesarios. Cuando el riesgo lo justifique, se aplicará una segmentación más granular, se supervisará el tráfico entre zonas y se probará que las reglas impiden desplazamientos laterales y accesos no autorizados sin comprometer la continuidad.

Justificación
La IA ofensiva puede acelerar el reconocimiento interno, el movimiento lateral y la propagación tras el compromiso de un componente. La segregación limita el alcance del incidente y dificulta el acceso a activos críticos desde zonas de menor confianza.

4.1.21 [mp.sw.1] – Desarrollo de aplicaciones

Código	Medida del ENS
[mp.sw.1]	**Desarrollo de aplicaciones**

Criterio de aplicabilidad
El desarrollo de aplicaciones controlará el código generado o asistido por inteligencia artificial, las dependencias introducidas automáticamente, la inyección de instrucciones procedente de fuentes no confiables, las acciones de agentes y la validación de entradas y salidas. Las mismas exigencias de revisión humana, pruebas y mínimo privilegio se aplicarán al software convencional y al que incorpore componentes de inteligencia artificial.

Justificación
La IA puede generar código inseguro, facilitar el descubrimiento y explotación de vulnerabilidades y ampliar la autonomía de los componentes.

4.1.22 [mp.sw.2] – Aceptación y puesta en servicio

Código	Medida del ENS
[mp.sw.2]	Aceptación y puesta en servicio

Criterio de aplicabilidad
La aceptación y puesta en servicio aportará evidencias de la resistencia de la aplicación a los escenarios de ataque potenciados mediante inteligencia artificial considerados en el análisis de riesgos y de las posibles vías de compromiso introducidas por código o componentes asistidos por inteligencia artificial, que se contrastarán con los criterios de aceptación y el riesgo residual admitido. Las pruebas incluirán, cuando proceda, análisis estático y dinámico, revisión de código, pruebas de penetración y escenarios automatizados o adaptativos.

Justificación
Los criterios de aceptación de seguridad se comprueban antes de la entrada en producción, en entornos adecuados.

4.1.23 [mp.s.1] – Protección del correo electrónico

Código	Medida del ENS
[mp.s.1]	Protección del correo electrónico

Criterio de aplicabilidad
Cuando así resulte del análisis de riesgos, la protección del correo electrónico reforzará la autenticación del correo, el análisis de contenido y comportamiento, la protección de URL y adjuntos y los procedimientos de verificación, de modo que mantenga su eficacia ante contenido sintético y campañas adaptativas.

Justificación
La IA aumenta la calidad y personalización del phishing, la suplantación y el compromiso de correo corporativo, y facilita la generación de mensajes, enlaces y adjuntos maliciosos.

4.1.24 [mp.s.2] – Protección de servicios y aplicaciones web

Código	Medida del ENS
[mp.s.2]	Protección de servicios y aplicaciones web

Criterio de aplicabilidad
Cuando así resulte del análisis de riesgos, la protección de servicios y aplicaciones web se llevará a cabo desplegando soluciones de gestión de tráfico automatizado capaces de diferenciar entre interacciones humanas legítimas y ataques automatizados apoyados o potenciados por herramientas basadas en inteligencia artificial empleadas con fines ofensivos, incluyendo, entre otros, actividades de extracción masiva de información, exploración sistemática de servicios e intentos de acceso no autorizado a cuentas.

Cuando así resulte del análisis de riesgos, se aplicarán además la validación de entradas, la limitación de tasas, la protección de API, el análisis de comportamiento y las pruebas frente a automatización.

Justificación
La IA puede incrementar la velocidad, escala y capacidad de adaptación de los ataques contra servicios web. La detección y contención se centran en el comportamiento malicioso, sin necesidad de acreditar el empleo de IA.

4.1.25 [mp.s.4] – Protección frente a denegación de servicio

Código	Medida del ENS
[mp.s.4]	**Protección frente a denegación de servicio**

Criterio de aplicabilidad
Las tecnologías desplegadas conforme a **[mp.s.4.2]** mitigarán incrementos anómalos o abusivos de tráfico, incluidos aquellos generados mediante herramientas de automatización o técnicas de inteligencia artificial empleadas con fines ofensivos. Se revisarán la capacidad, los límites, la redundancia, la protección anti-DDoS y la coordinación con proveedores.

Justificación
La IA puede facilitar el reconocimiento, la coordinación y la variación del tráfico ofensivo, por lo que conviene evaluar si aumenta su volumen, distribución o capacidad de evasión.

4.2 Matriz de escenarios y medidas relacionadas

La siguiente matriz facilita una propuesta de trazabilidad entre escenarios de riesgo y medidas de seguridad. **No sustituye al análisis de riesgos ni constituye una lista cerrada.**

La medida **[op.pl.1] Análisis de riesgos** abarca todos los escenarios aplicables al sistema, que deberán considerarse al identificar amenazas y evaluar y revisar los riesgos, incluido el residual. Para cada escenario, la matriz recoge las medidas más directamente vinculadas a su prevención, detección, respuesta o recuperación.

También son transversales **[op.mon.3] Vigilancia**, que mantiene actualizado el conocimiento de las amenazas, técnicas y capacidades ofensivas de IA relevantes para el sistema, y **[mp.per.4] Formación**, que prepara al personal técnico o de seguridad para identificarlas, prevenirlas y responder a ellas. Su aplicación y refuerzo se determinarán según la categoría del sistema, el análisis de riesgos y las funciones del personal.

Escenario	Descripción	Medidas relacionadas
Phishing y suplantación mediante IA	Uso de IA para generar comunicaciones, contenidos o interacciones altamente convincentes destinadas a suplantar identidades y obtener credenciales, información o la ejecución de acciones indebidas	[op.acc.3] Segregación de funciones y tareas [op.acc.5] Mecanismos de autenticación (usuarios externos) [op.acc.6] Mecanismos de autenticación (usuarios de la organización) [mp.per.3] Concienciación

Escenario	Descripción	Medidas relacionadas
		[mp.per.4] Formación [mp.s.1] Protección del correo electrónico
Explotación automatizada de vulnerabilidades	Empleo de IA para identificar, priorizar y explotar vulnerabilidades de forma automatizada, reduciendo el tiempo entre su descubrimiento y su aprovechamiento malicioso	[op.exp.2] Configuración de seguridad [op.exp.4] Mantenimiento y actualizaciones de seguridad [op.mon.1] Detección de intrusión [mp.s.2] Protección de servicios y aplicaciones web
Código dañino, generado o adaptado	Uso de IA para generar, modificar, depurar, traducir, ofuscar o mutar código dañino, producir variantes a mayor velocidad o adaptar artefactos para eludir mecanismos de detección y respuesta.	[op.exp.6] Protección frente a código dañino [op.mon.1] Detección de intrusión [op.exp.7] Gestión de incidentes [mp.com.4] Separación de flujos de información en la red
Inyección de instrucciones contra agentes de IA	Introducción de instrucciones maliciosas, de forma directa o a través de fuentes no confiables, destinada a alterar el comportamiento previsto de un agente o inducirle a realizar acciones no autorizadas.	[op.acc.4] Proceso de gestión de derechos de acceso [op.exp.2] Configuración de seguridad [op.exp.8] Registro de la actividad [mp.sw.1] Desarrollo de aplicaciones [mp.sw.2] Aceptación y puesta en servicio
Manipulación de salidas de IA	Generación o alteración de salidas destinadas a provocar efectos no deseados cuando son procesadas por otros componentes, aplicaciones, API o herramientas sin una validación suficiente.	[op.exp.2] Configuración de seguridad [mp.sw.1] Desarrollo de aplicaciones [mp.sw.2] Aceptación y puesta en servicio
Envenenamiento de fuentes RAG	Manipulación o incorporación de contenido malicioso en las fuentes de conocimiento utilizadas por un sistema RAG con el fin de condicionar sus respuestas o actuaciones.	[op.exp.1] Inventario de activos [op.ext.3] Protección de la cadena de suministro [mp.sw.1] Desarrollo de aplicaciones [mp.sw.2] Aceptación y puesta en servicio
Abuso de permisos o agencia excesiva	Utilización indebida por un agente de IA, o a través de este, de las herramientas, credenciales o permisos asignados para realizar acciones no autorizadas o de alcance superior al necesario.	[op.acc.3] Segregación de funciones y tareas [op.acc.4] Proceso de gestión de derechos de acceso [op.exp.8] Registro de la actividad [mp.com.4] Separación de flujos de información en la red
Exfiltración de información mediante IA	Obtención o transferencia no autorizada de información accesible al modelo o agente, incluidos datos, credenciales, secretos, contexto, instrucciones internas o información recuperada de otros sistemas.	[op.acc.4] Proceso de gestión de derechos de acceso [op.exp.2] Configuración de seguridad [op.exp.8] Registro de la actividad [mp.sw.1] Desarrollo de aplicaciones [mp.com.4] Separación de flujos de información en la red

Escenario	Descripción	Medidas relacionadas
Compromiso de la cadena de suministro	Introducción o aprovechamiento de componentes, modelos, datos, bibliotecas, servicios, API o actualizaciones comprometidos para afectar al sistema o facilitar accesos, manipulaciones o actuaciones no autorizadas	[op.exp.1] Inventario de activos [op.ext.3] Protección de la cadena de suministro [mp.sw.2] Aceptación y puesta en servicio
Consumo abusivo y denegación económica	Uso automatizado o inducido de modelos, agentes, API o herramientas para agotar capacidad, cuotas o presupuesto mediante solicitudes masivas, llamadas recursivas, bucles o tareas de coste elevado.	[op.exp.2] Configuración de seguridad [op.mon.1] Detección de intrusión [mp.s.4] Protección frente a denegación de servicio
Compromiso simultáneo o multietapa	Encadenamiento automatizado o coordinado de varias técnicas y fases de ataque, o afectación simultánea de múltiples identidades, servicios o dependencias, reduciendo el tiempo disponible para detectar, contener y recuperar.	[op.exp.7] Gestión de incidentes [op.cont.1] Análisis de impacto [op.cont.2] Plan de continuidad [op.cont.3] Pruebas periódicas [mp.com.4] Separación de flujos de información en la red

5 PRIORIZACIÓN DE LAS ACTUACIONES

Una vez determinada la aplicabilidad, la entidad priorizará las actuaciones teniendo en cuenta el riesgo que reducen y no por su número, de forma proporcionada a la categoría del sistema y a los resultados del análisis de riesgos, valorando al menos los siguientes criterios:

- **Exposición de activos y servicios**, explotación activa y posibilidad de automatización.
- **Impacto** sobre las dimensiones de seguridad y la criticidad de la información y de los servicios afectados.
- **Capacidad de agentes** o componentes para ejecutar acciones y alcances de sus permisos.
- **Dependencias de terceros,** concentración y dificultad de sustitución.
- **Comprensión del tiempo** disponible para detectar, decidir, contener y recuperar.
- **Eficacia,** urgencia, viabilidad, coste y plazo de implantación.
- **Riesgo residual y las medidas** compensatorias necesarias mientras se implantan las actuaciones.

Tendrán prioridad las actuaciones que reduzcan riesgos críticos, explotados activamente o capaces de causar efectos rápidos y amplios. El coste o la dificultad no justificarán por sí solos mantener sin tratamiento un riesgo no aceptado.

6 DOCUMENTACIÓN, SEGUIMIENTO Y REVISIÓN

Las decisiones de aplicabilidad, adaptación, prioridad y aceptación se justificarán en el análisis de riesgos y se reflejarán, cuando corresponda, en la **Declaración de Aplicabilidad y en los planes de tratamiento del riesgo.** Se documentarán responsables, plazos, evidencias, dependencias, excepciones, medidas compensatorias y riesgo residual.

La eficacia de las actuaciones se comprobará mediante registros, pruebas y ejercicios. Las decisiones se revisarán ante cambios relevantes en amenazas, exposición, arquitectura, permisos, componentes de IA, fuentes de conocimiento, proveedores, incidentes o los resultados de las verificaciones.