CCN-CERT

BP/36

Guía de buenas prácticas frente al modelo de IA ofensiva

GUÍA DE SEGURIDAD

JUNIO 2026

ccn-cert
centro criptológico nacional

Edita:

Fecha de Edición: junio de 2026

Índice

1. Sobre CCN-CERT, CERT Gubernamental Nacional

El CCN-CERT es la Capacidad de Respuesta a incidentes de Seguridad de la Información del Centro Criptológico Nacional, CCN, adscrito al Centro Nacional de Inteligencia, CNI. Este servicio se creó en el año 2006 como CERT Gubernamental Nacional español y sus funciones quedan recogidas en la Ley 11/2002 reguladora del CNI, el RD 421/2004 de regulación del CCN y en el RD 311/2022, de 3 de mayo, regulador del Esquema Nacional de Seguridad (ENS).

Su misión, por tanto, es contribuir a la mejora de la ciberseguridad española, siendo el centro de alerta y respuesta nacional que coopere y ayude a responder de forma rápida y eficiente a los ciberataques y a afrontar de forma activa las ciberamenazas, incluyendo la coordinación a nivel público estatal de las distintas Capacidades de Respuesta a Incidentes o Centros de Operaciones de Ciberseguridad existentes.

Todo ello, con el fin último de conseguir un ciberespacio más seguro y confiable, preservando la información clasificada (tal y como recoge el artículo 4.f de la Ley 11/2002 de 6 de mayo, reguladora del Centro Nacional de Inteligencia) y la información sensible, defendiendo el Patrimonio Tecnológico español, formando al personal experto, aplicando políticas y procedimientos de seguridad y empleando y desarrollando las tecnologías más adecuadas a este fin.

De acuerdo con esta normativa y la Ley 40/2015 de Régimen Jurídico del Sector Público es competencia del CCN-CERT la gestión de ciberincidentes que afecten a cualquier organismo o empresa pública. En el caso de operadores críticos del sector público la gestión de ciberincidentes se realizará por el CCN-CERT en coordinación con el CNPIC.

2. IA ofensiva: un cambio de paradigma que obliga a repensar la ciberseguridad

La incorporación de la inteligencia artificial al ámbito ofensivo está redefiniendo los fundamentos de la ciberseguridad. Su utilización ha dejado de ser una amenaza emergente para convertirse ya en una capacidad operativa integrada en campañas reales de actores criminales y estatales. Los casos documentados por la industria y por informes de inteligencia muestran fraudes multimillonarios apoyados en *deepfakes*, generación automatizada de *exploits* y campañas masivas altamente personalizadas.

Su impacto no reside únicamente en la aparición de nuevas amenazas, sino en **la capacidad de acelerar, escalar y automatizar técnicas ya conocidas, reduciendo drásticamente los tiempos** entre la identificación de una vulnerabilidad y su explotación. La automatización permite pasar a velocidad de máquina de la detección de una oportunidad al ataque efectivo, lo que reduce el margen de actuación defensiva y traslada gran parte de la ventaja operativa al atacante.

Las implicaciones para el **sector público** son especialmente relevantes por la criticidad de los servicios esenciales, la sensibilidad de la información gestionada, la interdependencia entre organismos y proveedores y la creciente conexión entre entornos IT y OT. En este contexto, la automatización ofensiva puede afectar a la

continuidad de servicios, incrementar la exposición de sistemas críticos y exigir una mayor coordinación institucional, intercambio de información, trazabilidad y control sobre los sistemas de IA desplegados.

Ante este cambio de paradigma, la adaptación debe apoyarse en unos fundamentos claros:

- **Volver a lo básico:** reforzar controles esenciales como la gestión de identidades, la segmentación, la monitorización continua, el control de accesos y la protección de activos críticos.
- **Transforma los procesos IT y OT:** integrar la seguridad desde el diseño, acelerar la gestión de vulnerabilidades y parcheo, automatizar validaciones y adaptar los controles a entornos operativos con baja capacidad de actualización.
- **Diseñar sistemas resilientes y seguros por defecto:** reducir la superficie de ataque, controlar dependencias, limitar conectividad innecesaria y preparar los sistemas para operar bajo condiciones adversas.
- **Usar la IA como actor defensivo gobernado:** incorporar capacidades de detección y respuesta automatizada bajo criterios estrictos de control, supervisión humana y trazabilidad.

La respuesta no debe limitarse a incorporar nuevas herramientas, sino que requiere revisar el modelo de seguridad en su conjunto. Esto implica integrar marcos de referencia de ciberseguridad y gobernanza de IA, alinear cumplimiento, seguridad técnica y control operativo, y establecer mecanismos de evaluación continua sobre sistemas, datos, modelos y agentes. La **IA defensiva** debe desplegarse con criterios de seguridad desde el diseño, control de permisos, protección de flujos de datos, monitorización permanente y revisión periódica del modelo de riesgo.

En definitiva, la IA ofensiva no representa una evolución incremental, sino un cambio de paradigma que exige avanzar hacia una ciberseguridad más continua, automatizada, resiliente y gobernada. Solo aquellas organizaciones capaces de combinar control, automatización y adaptación continua podrán reducir la asimetría frente a atacantes que ya operan con capacidades cada vez más autónomas, escalables y difíciles de anticipar.

3. Panorama del modelo de IA ofensiva

La **inteligencia artificial ofensiva** puede entenderse como el 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.

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 que pueden ejecutar sus acciones sin demasiados conocimientos. De hecho, **su principal cambio reside en la compresión de los tiempos de ataque**. La IA permite descubrir, encadenar y explotar vulnerabilidades en plazos mucho más reducidos, lo que limita la capacidad de reacción de organizaciones que todavía operan con ciclos de análisis, priorización, parcheo o respuesta pensados para semanas, no para horas.

Diversas fuentes apuntan esta dirección. Según Gartner, hace un año los agentes de IA reproducían vulnerabilidades reales con tasas inferiores al 10-30 %. Hoy, los modelos están cerca del **90 %** con escalada de intentos, descubren amenazas de día cero de forma autónoma y permiten descubrir y explotar vulnerabilidades sin conocimientos avanzados. Antes, la falta de conocimiento en los atacantes era una barrera adicional que protegía indirectamente. La IA reduce al mínimo esa protección haciendo que, con su apoyo, cualquiera pueda ser un atacante avanzado, lo que **generaliza estas capacidades y amplía significativamente el número de actores que pueden ejecutar ataques. [1][2][3]**

El caso de **PROMPTSPY**, muestra como el *malware* puede integrar IA generativa en su flujo de ejecución para tomar decisiones dinámicas en función del entorno. Este *malware* utiliza modelos como *Gemini* para interpretar el estado del sistema en tiempo

real, generar instrucciones de acción paso a paso y adaptarse automáticamente a diferentes dispositivos o configuraciones. **[4]**

Por su parte, investigaciones del ***Institute for AI Policy and Strategy*** **[5]** documentan campañas reales donde agentes IA ejecutaban **entre el 80% y el 90% de las operaciones ofensivas**, dejando al humano en un rol supervisor.

En la misma línea, organismos como NIST y CISA advierten que la IA incrementa de forma significativa la **velocidad, escala y sofisticación de los ataques**, afectando a todas las fases del ciclo de vida y exigiendo nuevos enfoques de gestión del riesgo **[6]**. También ENISA se ocupó en 2023 de hacer una aproximación a los desafíos que supone la IA mediante la publicación del documento *Artificial Intelligence and Cybersecurity Research* **[7]**.

La inteligencia artificial se ha consolidado como una **prioridad estratégica** para las organizaciones, con un crecimiento acelerado tanto en inversión como en despliegue operativo. Según *Deloitte* **[8]**, alrededor del **60 % de los empleados ya tiene acceso a herramientas de IA**, lo que refleja una adopción masiva en un corto periodo de tiempo. Sin embargo, este incremento en el uso no se traduce de forma directa en madurez: solo una parte limitada de las organizaciones ha logrado pasar de la experimentación a la industrialización, y apenas el 34 % está transformando realmente su forma de operar con IA. En este contexto, la IA agéntica emerge como la próxima fase de evolución, con tasas de adopción previstas que podrían alcanzar el 74 % en los próximos dos años, frente al 23 % actual. No obstante, esta rápida expansión se produce en un escenario de riesgo significativo, ya que **sólo** aproximadamente **una de cada cinco organizaciones dispone de modelos maduros de gobernanza y control sobre estos sistemas autónomos. [5]** Esta combinación de alta adopción, impacto creciente y baja madurez en control sitúa a la IA agéntica como un factor clave de oportunidad y, al mismo tiempo, como **uno de los principales riesgos emergentes en ciberseguridad**, especialmente en lo relativo a la gestión de identidades, el acceso a datos y la ejecución autónoma de acciones críticas.

La IA ofensiva no introduce amenazas nuevas: colapsa los plazos de las que ya conocemos. Lo que protegía ayer ya no protege mañana, porque el atacante ahora opera a velocidad de máquina, en paralelo y sin descansar. **La respuesta no debe limitarse a la adquisición de nuevas herramientas, exige rediseñar los procesos [4].**

El atacante con IA descubre, encadena y explota vulnerabilidades en tiempos muy reducidos. La higiene básica deja de ser una buena práctica y pasa a ser la única línea de contención y defensa.

Los procesos de gestión de cambios, de parcheo y de compras fueron diseñados para ciclos de semanas y con una metodología tediosa de preparación y acción. La IA ofensiva opera en horas. **Si los procesos no cambian, las herramientas no llegan a tiempo. [3]**

La asimetría se cierra sólo si el equipo de seguridad opera con IA. Pero esos agentes son, a su vez, una nueva superficie de ataque que hay que gobernar como cualquier sistema crítico. **[5]**

3. Panorama del modelo de IA ofensiva

[1] *Gartner*, https://www.gartner.com/en/documents/6579402

[2] *Thehackernews*, https://thehackernews.com/2026/05/hackers-used-ai-to-develop-first-known.html

[3] *Securityweek*, https://www.securityweek.com/the-zero-knowledge-threat-actor-and-the-end-of-responsible-disclosure/

[4] *Google*, https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools/
https://blog.google/innovation-and-ai/infrastructure-and-cloud/google-cloud/gtig-report-ai-cyber-attacks-feb-2026/

[5] *Institute for AI Policy and Strategy*, https://www.iaps.ai/

[6] *NIST - CISA*, https://www.nist.gov/itl/ai-risk-management-framework
https://www.cisa.gov/ai

[7] *ENISA*, https://www.enisa.europa.eu/publications/artificial-intelligence-and-cybersecurity-research

[8] *Deloitte*, https://www.deloitte.com/es/es/services/consulting/research/estado-ia-en-las-empresas.html
https://www.deloitte.com/content/dam/assets-zone2/es/es/docs/services/consulting/2026/state-of-ai-2026.pdf
https://www.deloitte.com/es/es/services/consulting/research/estado-ia-generativa-empresas-espana.html
https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-ai-in-the-enterprise.html
https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/articles/agentic-ai-insights.html

3.1. Contexto actual de la defensa mediante IA

Uso inseguro de la IA en las organizaciones

La IA está siendo adoptada por las líneas de negocio más rápido de lo que maduran los controles de seguridad, creando nuevas superficies de ataque. Este desajuste genera riesgos asociados al uso no gobernado de herramientas, modelos y agentes, especialmente cuando se incorporan a procesos internos sin criterios claros de validación, trazabilidad o control de acceso.

Por ello, los responsables de ciberseguridad deben implantar un modelo de **validación específica para sistemas de inteligencia artificial**, basado en estándares reconocidos que permitan identificar riesgos propios de estos entornos. Utilizar referencias como OWASP para aplicaciones de IA y modelos de lenguaje, MITRE ATLAS para simular técnicas de ataque y guías de NIST y ENISA para evaluación continua y gestión del riesgo.

Este enfoque debe materializarse en pruebas de penetración adaptadas a IA, ejercicios de simulación de ataques y validación recurrente de modelos y agentes, integrados en el ciclo de vida. En este sentido, **LabIA,[1] la plataforma de retos de inteligencia artificial del Centro Criptológico Nacional** es una herramienta muy adecuada para fortalecer las capacidades en seguridad aplicada a modelos de IA.

Alfabetización en IA

Los responsables de ciberseguridad deben desarrollar un conocimiento base sobre los nuevos riesgos introducidos por la inteligencia artificial, apoyándose en marcos regulatorios y guías oficiales que permitan comprender su impacto real. En este sentido, resulta clave utilizar como referencia el **Reglamento Europeo de Inteligencia Artificial *(AI Act)***, que establece obligaciones en gestión de riesgos, gobernanza y supervisión de sistemas de IA, así como apoyarse en estándares como el ***AI Risk Management Framework del NIST*** y las recomendaciones de **ENISA sobre ciberseguridad en IA**, que aportan criterios prácticos para la identificación, evaluación y mitigación de amenazas emergentes. Por otra parte, todo lo anterior se debe entender en el marco más amplio de las obligaciones legales de gestión del riesgo tecnológico que establecen principalmente dos normas: la Directiva NIS 2 y el Reglamento DORA. A ellas hay que sumar el RGPD, ya que en muchas ocasiones los usuarios comparten datos de carácter personal con los modelos.

Gartner **[2]** propone la figura del Experto en IA dentro de los equipos de seguridad para consolidar este conocimiento. En el sector privado se está consolidando la aparición de una figura ejecutiva específica para liderar la transformación

asociada a la inteligencia artificial, conocida como ***Chief Artificial Intelligence Officer (CAIO)***, concebida como un rol equivalente en relevancia al CISO pero centrado en la **estrategia, gobernanza y riesgos de la IA**. Esta posición surge como respuesta a la creciente complejidad de la adopción de IA y a la necesidad de centralizar su control, ya que las organizaciones están trasladando la responsabilidad de estos sistemas al nivel del comité ejecutivo, con funciones que abarcan desde la definición estratégica hasta la supervisión del uso seguro y responsable de la tecnología **[3]**. En el Gobierno federal de los Estados Unidos, la OMB ha exigido a las agencias federales, mediante los memorandos M-24-10 y M-25-21, la designación de un Chief AI Officer como figura responsable de promover la adopción, gobernanza y gestión del riesgo de IA.

Herramientas de seguridad para IA

Están emergiendo nuevas tecnologías específicas que, aunque todavía son inmaduras, serán clave para descubrir y controlar el uso de la IA, incluyendo la *Shadow AI*.

Las capacidades esenciales pueden apoyarse en soluciones concretas:

- **Protección de interacción y uso de modelos** como *Microsoft Azure AI Security / Prompt Shields y Lakera Guard*, que permiten filtrar entradas, mitigar manipulaciones (como *prompt injection*) y controlar la salida de información sensible.
- **Seguridad del ciclo de vida de IA (modelos, datos y desarrollo)** como *Protect AI* y *HiddenLayer*, que facilitan la identificación de vulnerabilidades en modelos y procesos, así como la protección frente a manipulación y riesgos en la cadena de suministro.
- **Detección y respuesta operativa** como *Microsoft Sentinel, Elastic Security, CounterCraft*, que permiten monitorizar, detectar comportamientos anómalos y responder a incidentes, incluyendo técnicas avanzadas como *deception* para entornos críticos.
- **Gobierno, control y cumplimiento normativo** como *IBM watsonx. governance* y Marcos ENS, NIST AI RMF, AI Act, que facilitan el control del uso, la trazabilidad, la auditoría y el cumplimiento regulatorio.
- **Seguridad de sistemas autónomos y agentes** como *Alias Robotics*, especializada en protección de sistemas autónomos y agentes, relevante en escenarios de automatización avanzada.

Es importante empezar a planificar su integración en el conjunto tecnológico y en las operaciones del centro de operaciones de seguridad.

Gobernanza como base

La gobernanza de la IA se convertirá en un pilar para la ciberseguridad.

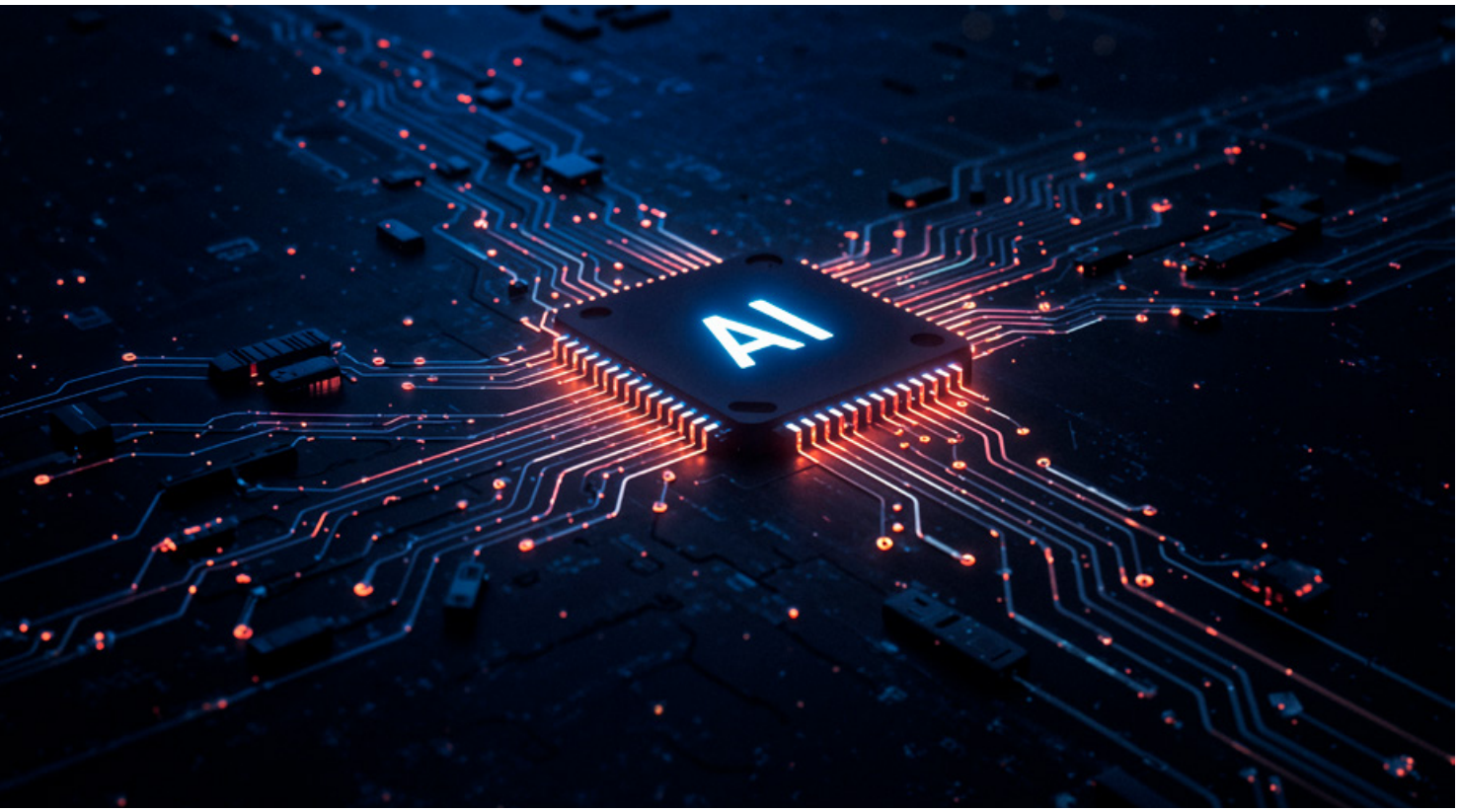
La gestión del riesgo asociado al uso de inteligencia artificial debe alinearse

con los marcos regulatorios y de gobernanza definidos a nivel europeo y nacional, en lugar de centrarse únicamente en enfoques de mercado o recomendaciones de fabricantes. En este sentido, la implantación de controles debe basarse en los principios establecidos por el **Reglamento Europeo de Inteligencia Artificial (AI Act),** que introduce obligaciones específicas en materia de gestión de riesgos, transparencia y supervisión humana, así como en las directrices de ciberseguridad definidas por organismos como **ENISA**, que destacan la necesidad de asegurar todo el ciclo de vida de los sistemas IA. A nivel nacional, estas recomendaciones deben integrarse con el **Esquema Nacional de Seguridad (ENS)** y las guías **STIC** del **CCN-CERT**, garantizando un enfoque coherente con la gestión global de la seguridad de la información. Este marco combinado permite estructurar la protección frente a amenazas avanzadas, incluyendo aquellas derivadas de la IA ofensiva, mediante modelos de control continuo, evaluación de riesgos dinámicos y supervisión efectiva de los sistemas desplegados.

[1] *LabIA*, https://labia.ccn.cni.es

[2] *Gartner*, https://www.gartner.com/en/documents/6579402

[3] *IBM*, https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/augmented-workforce



3.2. Amenazas, riesgos y posibles escenarios en el uso de la IA

La adopción de Inteligencia Artificial en el cibercrimen ha pasado de un escenario experimental a uno operativo y confirmado en incidentes reales. Informes de inteligencia y casos documentados muestran que la IA:

Reduce la complejidad técnica de los ataques.

- El informe del ***Center for Long-Term Cybersecurity (UC Berkeley)*** indica que los modelos generativos están **reduciendo las barreras de entrada**, permitiendo a actores con menor conocimiento técnico generar *phishing*, *malware* y tareas de reconocimiento con facilidad. **[1]**
- El informe de **ENISA *Threat Landscape*** 2025 confirma que la disponibilidad de herramientas como *FraudGPT* o *WormGPT* y el modelo "*phishing-as-a-service*" permite que **actores poco cualificados** ejecuten campañas avanzadas, expandiendo el ecosistema criminal. **[2]**
- Investigaciones académicas muestran que los ataques potenciados por IA **reducen la complejidad operativa hasta en un 72 %**, automatizando tareas que antes requerían alta especialización. **[3]**

Incrementa la tasa de éxito.

- El ***Microsoft Security Blog* (2026)** reporta que el uso de IA en campañas de *phishing* eleva las ratios de éxito hasta un **54 % frente** a aproximadamente **un 12 %** en campañas tradicionales (≈450 % de mejora). **[4]**
- Estudios académicos también confirman que los ataques asistidos por IA alcanzan **hasta un 67 % más de efectividad**, gracias a su capacidad de adaptación y personalización. **[5]**
- ENISA señala que el *phishing* asistido por IA es ya dominante y más preciso y efectivo, ya que permite adaptar mensajes a perfiles concretos y contextos reales. **[6]**

Permite automatizar campañas a gran escala.

- El ***Google Threat Intelligence Group* (GTIG)** documenta que actores estatales utilizan IA para **automatizar tareas de reconocimiento,**

generación de *phishing* y desarrollo de *malware* a escala, integrándola en todo el ciclo de ataque. **[7]**

- El informe de **ENISA 2025** indica que el *phishing* se está ejecutando a **"escala máquina"**, con más del **80 % de campañas incorporando IA** y siendo generadas de forma automatizada. **[8]**
- Informes de inteligencia y análisis global muestran que los atacantes utilizan IA para **automatizar escaneos masivos (hasta 36.000 intentos por segundo)** y ejecutar actividades de reconocimiento y ataque de forma continua. **[9]**, **[10]**

[1] https://cltc.berkeley.edu/2026/02/11/new-cltc-report-on-managing-risks-of-agentic-ai/

[2] https://digital4security.eu/enisa-threat-landscape-2025/

[3] https://pdfs.semanticscholar.org/

[4] https://www.microsoft.com/en-us/security/blog/

[5] https://pdfs.semanticscholar.org/

[6] https://socket.dev/blog/enisa-threat-landscape-2025

[7] https://blog.google/threat-analysis-group/google-tag-generative-ai/

[8] https://www.cybersixt.com/enisa-threat-landscape-2025

[9] https://www.fortinet.com/blog/threat-research

[10] https://www.allaboutai.com/cybersecurity/ai-attacks/

Amenazas identificadas

- ***Phishing* avanzado impulsado por IA:** Evolución hacia campañas masivas muy dirigidas y adaptadas más aún al objetivo, generadas automáticamente y capaces de escalar con alta efectividad, incrementando significativamente el volumen y superando los mecanismos tradicionales de detección.

- **Suplantación avanzada mediante *deepfakes*:** Uso de IA para replicar voz, imagen y comportamiento de personas reales, permitiendo ataques en tiempo real con alto impacto económico.

- **Generación de *malware* y código malicioso (*exploits*) mediante IA:** La IA se utiliza para desarrollar código malicioso y descubrir vulnerabilidades. Por ejemplo, el equipo de ciberseguridad de Google identificó el primer *exploit* de día cero desarrollado con ayuda de IA, capaz de evadir autenticación de segundo

factor. Este *exploit* fue diseñado para una operación de explotación masiva. Se detectaron características típicas de código generado por modelos LLM.

- **Automatización del reconocimiento en fuentes abiertas:** Uso de IA para recopilar y correlacionar información pública y técnica. La IA se usa para realizar de manera más rápida análisis de datos robados, perfilado automático de víctimas e identificación masiva de objetivos.

  En investigaciones de *Google Threat Intelligence Group* **[1]**, se ha observado que los atacantes utilizan IA como un "**asistente de investigación a gran escala**" para recopilar información sobre objetivos, identificar tecnologías usadas (conjunto tecnológico, servicios expuestos, versiones) y analizar rápidamente posible superficie de ataque.

  Un ejemplo es el uso de IA por actores avanzados para automatizar el reconocimiento de objetivos antes del ataque, combinando fuentes públicas con análisis automatizado. Actores estatales y grupos organizados utilizan modelos de IA para analizar grandes volúmenes de información pública (redes sociales, webs corporativas, repositorios) y generar perfiles detallados de víctimas e identificar organizaciones vulnerables de forma masiva.

  Mediante la combinación de **fuentes abiertas de información (OSINT)** con capacidades de **reconocimiento directo sobre las infraestructuras**, los atacantes pueden mapear de forma precisa activos digitales como sitios web, interfaces de programación, servicios en la nube o entornos de infraestructura como servicio. La IA permite además **correlacionar automáticamente estos datos**, identificar tecnologías utilizadas, detectar configuraciones vulnerables y priorizar objetivos con mayor probabilidad de éxito. El reconocimiento deja de ser una fase preliminar limitada y se convierte en un proceso dinámico y permanente que incrementa significativamente la **velocidad, precisión y alcance de los ataques**, reduciendo el esfuerzo necesario y ampliando la exposición real de las organizaciones.

- **Ingeniería social automatizada y adaptativa:** La automatización mediante IA permite lanzar millones de ataques altamente creíbles en poco tiempo, convirtiendo este vector en el principal mecanismo de acceso inicial. La ingeniería social automatizada y adaptativa se ve potenciada por la IA, ya que permite crear canales de conversación automatizados (*chatbots*) maliciosos, generar respuestas dinámicas personalizadas y producir simulaciones humanas convincentes, lo que facilita que los ataques se ajusten al contexto y a la interacción con la víctima en tiempo real.

- **Ataques a sistemas IA:** son ataques dirigidos a manipular sistemas basados en modelos de lenguaje. Entre sus características se incluyen el secuestro del comportamiento del modelo, la exfiltración de información y la ejecución de acciones no autorizadas:

- Exfiltrar información de contexto.
- Evadir restricciones del sistema.
- Manipular decisiones automáticas.
- Alterar procesos de análisis documental o RAG (***Retrieval-Augmented Generation*** - técnica de IA que conecta un modelo de lenguaje con fuentes de datos privadas o externas).
- Influir en agentes autónomos que tengan acceso a herramientas.

Casos documentados como el de ***Microsoft 365 Copilot*** **[2]** o ***Slack AI*** **[3]** demuestran que un atacante puede introducir instrucciones ocultas en correos, documentos o mensajes aparentemente legítimos, logrando que el sistema ignore sus reglas originales y ejecute acciones no previstas, como acceder a información interna o filtrarla hacia destinos externos sin conocimiento del usuario.

Este tipo de ataques aprovecha una debilidad estructural de los modelos, incapaces de diferenciar de forma fiable entre datos e instrucciones, lo que permite secuestrar el comportamiento del sistema, exfiltrar información sensible e incluso desencadenar acciones automatizadas sobre sistemas conectados. En consecuencia, los LLM dejan de ser únicamente una superficie de exposición informativa para convertirse en un punto crítico de control operativo dentro de la infraestructura, con implicaciones directas en la seguridad organizativa.

A estos vectores se añade un riesgo estructural propio de los sistemas con memoria persistente o capacidades de reentrenamiento: la acumulación progresiva de contexto entre sesiones. A diferencia de los modelos sin memoria, estos sistemas retienen información de interacciones previas que puede combinarse con nuevos datos para generar conexiones no previstas entre conocimientos aparentemente inconexos. Un adversario que interactúe de forma deliberada y continuada con estos sistemas puede explotar ese conocimiento acumulado para identificar patrones, inferir información sensible o encadenar acciones que el defensor —humano o automatizado— no había anticipado. Este riesgo es especialmente relevante en arquitecturas con acceso a bases de conocimiento externas —como arquitecturas RAG— o en entornos donde múltiples agentes comparten contexto. **[4]**

La IA agéntica introduce una transformación profunda en la ciberseguridad, tanto desde el punto de vista ofensivo como defensivo. Por un lado, permite automatizar el ciclo completo del ataque, desde el reconocimiento hasta la exfiltración, mediante análisis de datos, toma de decisiones contextual y ejecución autónoma, lo que incrementa la velocidad, la escala y la adaptabilidad, reduciendo la necesidad de intervención humana. Además, la incorporación de estos sistemas como "trabajadores digitales" abre nuevos vectores de riesgo, al poder ser manipulados o utilizados para acceder a información crítica o ejecutar acciones no autorizadas.

- Por otro lado, estas mismas capacidades habilitan una evolución hacia modelos de **ciberseguridad más automatizados y proactivos**, donde los agentes permiten detectar anomalías, analizar eventos, coordinar respuestas y gestionar el riesgo en tiempo real, mejorando la eficiencia operativa y reduciendo los tiempos de reacción. Sin embargo, esta dualidad genera un reto clave: la **brecha entre adopción y gobernanza**, ya que la implantación de estos sistemas avanza a mayor velocidad que los mecanismos de control, lo que introduce riesgos en términos de seguridad, privacidad y trazabilidad.

- En este contexto, la clave reside en avanzar hacia un modelo de **autonomía gobernada**, basado en control, supervisión y limitación de capacidades, asegurando visibilidad sobre el comportamiento de los agentes y manteniendo intervención humana en decisiones críticas. En definitiva, la IA agéntica redefine el equilibrio entre atacante y defensor, situando la ventaja en aquellos que sean capaces de operar estos sistemas con mayor control, visibilidad y velocidad.

La evolución reciente de la inteligencia artificial hacia modelos agénticos (sistemas capaces de percibir, razonar, decidir y ejecutar acciones de forma autónoma) está introduciendo un cambio estructural en el ámbito de la ciberseguridad. Según los informes de Deloitte sobre el estado de la IA, las organizaciones están acelerando la adopción de estos sistemas, pasando rápidamente de fases piloto a entornos productivos, lo que amplía significativamente la superficie de riesgo si no se dispone de mecanismos adecuados de control y supervisión. En este contexto, la aparición de agentes avanzados como OpenClaw o Hermes evidencia un cambio relevante en la capacidad de orquestación, especialmente cuando se combinan con modelos locales sin restricciones, a los que se han eliminado los mecanismos de seguridad y que responden íntegramente según su entrenamiento. Esta evolución facilita la automatización de tareas tradicionalmente complejas como el escaneo de redes, la identificación de vulnerabilidades o incluso la explotación, permitiendo ejecutar de forma autónoma o semiautónoma toda la cadena de ataque, con una reducción significativa del esfuerzo humano y un incremento sustancial en la velocidad y escala de las operaciones.

---

[1] *Google,* https://blog.google/threat-analysis-group/google-tag-generative-ai/
https://blog.google/threat-analysis-group/updates-on-threat-actors-use-of-ai/
https://blog.google/threat-analysis-group/new-report-on-ai-and-security/

[2] *Microsoft 365 Copilot, https://arxiv.org/abs/2509.10540*
https://nvd.nist.gov/vuln/detail/CVE-2025-32711
https://thehackernews.com/2025/06/zero-click-ai-vulnerability-exposes.html

[3] *Slack AI, https://www.theregister.com/software/2024/08/21/slack-ai-can-leak-private-data-via-prompt-injection/1073730*
https://www.darkreading.com/cyberattacks-data-breaches/slack-ai-patches-bug-that-let-attackers-steal-data-from-private-channels
https://cybersecuritynews.com/critical-slack-vulnerability/
https://www.startupdefense.io/mitre-atlas-case-studies/aml-cs0035-data-exfiltration-from-slack-ai-via-indirect-prompt-injection

[4] *AEPD, https://www.aepd.es/guias/orientaciones-ia-agentica.pdf*

Relación de riesgos, ejemplos y controles

FAMILIA DE RIESGO	DESCRIPCIÓN	EJEMPLOS	CONTROLES RECOMENDADOS
Automatización ofensiva	Uso de IA para acelerar reconocimiento, explotación y movimiento lateral.	Reconocimiento masivo, explotación asistida, generación de *payloads*.	CTEM, ASM, KEV/EPSS, detección basada en comportamiento.
Ingeniería social sintética	Uso de IA para suplantación, *phishing*, voz o vídeo falso.	*Phishing* dirigido, *deepfake*, clonación de voz.	Verificación fuera de banda, formación, controles antifraude.
Ataques contra sistemas IA	Manipulación de modelos, agentes o flujos RAG.	*Prompt injection,* filtración de datos, abuso de la herramienta.	Validación de entradas/ salidas, aislamiento, auditoría.
Abuso de agentes autónomos	Agentes que ejecutan acciones sin control suficiente.	Llamadas API, cambios, extracción de datos.	Mínimo privilegio, supervisión humana, *kill switch*.
Riesgo de cadena de suministro IA	Dependencias, modelos, conectores, complementos, MCP, terceros.	Componentes no gobernados, *shadow AI*.	Inventario IA, SBOM/AI-BOM, evaluación de terceros.

Posibles escenarios

- **Explotación acelerada de vulnerabilidades expuestas a Internet**
  Un atacante detecta una vulnerabilidad publicada en un servicio web y la explota en pocas horas antes de que la organización pueda aplicar el parche.
- ***Phishing* hiperpersonalizado contra directivos o personal con privilegios**
  Se envían correos generados por IA con información real del destinatario para inducirle a aprobar accesos o transferencias fraudulentas.
- ***Deepfake* de voz o vídeo para autorización fraudulenta**
  Se utiliza una videollamada o mensaje de voz simulado de un directivo para autorizar pagos o cambios críticos.

- **Uso de IA para reconocimiento masivo de organismos públicos**
  Un sistema automatizado analiza webs, documentos y servicios expuestos para identificar vulnerabilidades y puntos de acceso.

- **Agente IA comprometido que ejecuta acciones no autorizadas**
  Un agente integrado en procesos internos es manipulado para acceder a datos sensibles o realizar operaciones fuera de su función.

- **Proveedor afectado por vulnerabilidad descubierta y explotada con IA**
  Un fallo en un proveedor tecnológico es identificado automáticamente y explotado, afectando en cascada a múltiples organismos dependientes.

- **Fuga de información mediante *prompt injection* o abuso de conectores**
  Un *input* malicioso hace que un sistema IA acceda a información interna y la exponga en sus respuestas o la envíe a terceros.

3.3. Implicaciones estratégicas para el sector público

La evolución reciente de la inteligencia artificial aplicada a la ciberseguridad, especialmente con la aparición de modelos avanzados como *Claude Mythos Preview* o *GPT-5.5-Cyber*, introduce un cambio estructural en el equilibrio entre atacantes y defensores. Este cambio afecta de forma directa a la seguridad, resiliencia y gobernanza del sector público, transformando la ciberseguridad en un problema de velocidad, escala y capacidad de adaptación continua.

El uso de inteligencia artificial permite ejecutar ataques a velocidad máquina, reduciendo drásticamente los tiempos de intrusión y explotación. Las intrusiones pueden progresar en minutos y los ciclos de ataque se han comprimido desde días o semanas a tiempos cercanos a la ejecución inmediata, reduciendo la capacidad de respuesta de las organizaciones.

- **Colapso del ciclo de vulnerabilidades**

  La inteligencia artificial ha transformado el ciclo tradicional de vulnerabilidades. Actualmente, la identificación y explotación pueden producirse casi simultáneamente, eliminando la ventana temporal en la que las organizaciones

podían reaccionar antes de ser comprometidas. Este fenómeno hace que los modelos clásicos de gestión de vulnerabilidades queden obsoletos.

La aparición de modelos avanzados de IA especializados en ciberseguridad representa un punto de inflexión. Estos modelos pueden identificar vulnerabilidades a gran escala, generar *exploits* complejos y encadenar ataques de forma autónoma. Su capacidad para encontrar miles de vulnerabilidades críticas en sistemas ampliamente utilizados evidencia un cambio cualitativo en el riesgo global.

Las vulnerabilidades ya no se limitan únicamente nuevos CVE o **vulnerabilidades de día cero**, también se amplía el abanico a **configuraciones inseguras, exposiciones indebidas y riesgos derivados de terceros**. La automatización del reconocimiento mediante IA permite identificar de forma sistemática estos puntos débiles a lo largo de toda la superficie de ataque, incluyendo servicios mal configurados, accesos y permisos excesivos o dependencias externas comprometidas. Como resultado, el foco de riesgo se amplía desde el código hacia el conjunto del entorno operativo, incrementando la probabilidad de exposición y la complejidad de su gestión. Todo lo anterior también tiene un impacto en materia de cumplimiento normativo, ya que el Reglamento UE 2024/2847 del Parlamento Europeo y del Consejo, de 23 de octubre de 2024, relativo a los requisitos horizontales de ciberseguridad para los productos con elementos digitales, conocido como *Cyber Resilience Act*, crea un nuevo régimen legal de gestión y notificación de vulnerabilidades que entra en vigor en septiembre de 2026.

- **Implicaciones a corto plazo:** A corto plazo se espera un aumento significativo en el número de vulnerabilidades detectadas y explotadas, un incremento de campañas automatizadas de ataque y una reducción de los tiempos de parcheo disponibles. Las administraciones públicas deberán enfrentarse a entornos donde la exposición se materializa casi inmediatamente tras la aparición de una vulnerabilidad.
- **Implicaciones a medio plazo:** A medio plazo, la industrialización del ciberataque permitirá que actores con menor capacidad técnica ejecuten ataques avanzados. La automatización, la escalabilidad y la reducción de costes operativos del atacante incrementarán el riesgo sistémico sobre infraestructuras críticas y servicios públicos.

Exposición del factor humano

La inteligencia artificial amplifica la eficacia de la ingeniería social mediante técnicas como *phishing* avanzado, *deepfakes* y generación de contenido personalizado. Esto reduce la capacidad del personal para identificar ataques, convirtiendo al factor humano en un vector de compromiso más crítico.

Impacto en la continuidad de servicios

La automatización y velocidad del ataque incrementan el riesgo de interrupciones en servicios públicos esenciales. La explotación simultánea de vulnerabilidades en sistemas interconectados puede generar fallos en cascada con impacto directo en la ciudadanía.

Impacto en gobernanza y regulación

La naturaleza dual de estos modelos obliga a replantear los marcos regulatorios y de control. La gestión de capacidades como Mythos se convierte en un asunto estratégico que afecta a la seguridad nacional, la estabilidad económica y la protección de infraestructuras críticas.

Cambio de paradigma en la ciberseguridad pública

El sector público debe evolucionar hacia modelos de seguridad continua, automatizada y basada en inteligencia. La integración de capacidades de detección en tiempo real, respuesta automatizada y uso de IA defensiva se convierte en un requisito esencial para mantener la resiliencia frente a amenazas que operan a velocidad máquina.

3.4. Fundamentos para la adaptación

>> La respuesta no debe limitarse a la adquisición de nuevas herramientas, exige rediseñar los procesos actuales para adaptarlos al ritmo que puede imprimir la automatización y agilidad de los agentes IA como atacantes.

VOLVER A LO BÁSICO

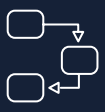
"Lo que ya sabíamos hacer, ahora hay que hacerlo a velocidad de máquina"

>> La ciberseguridad ya no es sólo un problema de conocimiento, sino de velocidad de ejecución, donde los controles deben operar en tiempo real para no quedar sistemáticamente por detrás del atacante.

El atacante con IA descubre, encadena y explota vulnerabilidades en horas. La higiene básica deja de ser una buena práctica y pasa a ser la única línea de contención.

VOLVER A LO BÁSICO

"La IA no modifica nuestra estrategia ciber, pero sí acelera la necesidad de fortalecer lo básico"



>> Las infraestructuras tecnológicas actuales están diseñadas para amenazas humanas, hay que rediseñarlas.

Los procesos de gestión de cambios, de parcheo, compras, etc., fueron diseñados para ciclos de semanas y con una metodología tediosa de preparación y acción. La IA ofensiva opera en horas. Si los procesos no cambian, las herramientas no llegan a tiempo.

CAMBIAR LOS PROCESOS DE IT

"No basta con implementar IA defensiva, hay que reformular el tiempo de los procesos que la rodean"

>> Acelerar procesos parece suficiente, pero no lo es. Incrementar la velocidad de escaneo o de parcheo no resuelve que los ciclos de validación, pruebas y despliegue sigan operando en tiempos incompatibles, y forzarlos implica riesgo y aumenta las probabilidades de introducir fallos mayores que los que se pretenden corregir. El problema no es solo de rapidez, es de diseño.

Las organizaciones deben asumir que siempre existirán vulnerabilidades y que **el objetivo debe ser dificultar su explotación desde la arquitectura**, mediante defensas en profundidad, aislamiento de entornos y despliegues centralizados. En este contexto, la ventaja competitiva ya no estará sólo en reaccionar más rápido, sino en construir sistemas resilientes por defecto, donde la seguridad no dependa únicamente del tiempo de respuesta, sino del propio diseño del entorno.

DISEÑO RESILIENTE Y SEGURO POR DEFECTO

"El futuro de la ciberseguridad no depende solo de ir más rápido, sino de rediseñar sistemas para ser resistentes por defecto"

>> Si el atacante usa agentes, el defensor tiene que usarlos también, pero de **forma controlada**.

La asimetría se cierra solo si el equipo de seguridad opera con IA. Pero esos agentes son, a su vez, una nueva superficie de ataque que hay que gobernar como cualquier sistema crítico.

LA IA COMO UN ACTOR DEFENSIVO

"De IA como riesgo a IA como herramienta de trabajo"

4. Buenas prácticas

4.1. Buenas prácticas de procesos IT

Procesos tradicionalmente estáticos, que deben adaptarse al nuevo paradigma: gestión de vulnerabilidades, gestión de cambios, ciclo de vida del desarrollo de *software* y seguridad de aplicaciones, identidad y accesos, copias de seguridad y recuperación ante desastres, y riesgo de terceros.

GESTIÓN DE VULNERABILIDADES

AHORA:

En el modelo actual el parcheo se planifica en ventanas temporales, se prioriza por CVSS y se evita tocar directamente producción salvo emergencia.

A DONDE TENEMOS QUE IR:

Debemos dirigirnos a un modelo en el que se realice un **escaneo y parcheo continuo**, priorizado según la **explotabilidad real** de las vulnerabilidades (KEV, EPSS). En este contexto, es más importante que nunca **identificar con precisión los activos críticos**, de modo que sean prioritarios en revisión y parcheo, independientemente de las capas de defensa existentes.

Asimismo, resulta clave habilitar mecanismos de **auto remediación para escenarios acotados**, ejecutando acciones correctivas de forma autónoma, pero siempre bajo un conjunto bien definido de casos de uso y con **controles de seguridad estrictos (guardarraíles)** que garanticen el control y eviten impactos no deseados.

GESTIÓN DE CAMBIOS

AHORA:

Cada cambio pasa por un proceso con formularios, aprobaciones manuales y ventanas de despliegue rígidas.

A DONDE TENEMOS QUE IR:

En el futuro los cambios viajan como código en *pipelines* automatizados. La decisión humana sólo entra a decidir sobre los cambios verdaderamente críticos.

Esto implica una documentación exhaustiva de los cambios, para poder realizar una vuelta atrás con conocimiento de los cambios introducidos.

CICLO DE VIDA DEL DESARROLLO DE SOFTWARE Y SEGURIDAD DE APLICACIONES

AHORA:

El modelo con el que se desarrolla revisa la seguridad una vez por despliegue, al final del ciclo, y el equipo de desarrollo la vive como un trámite.

A DONDE TENEMOS QUE IR:

Cada cambio propuesto en el código debe ser analizado automáticamente por la IA para buscar puntos débiles, un analizador revisa el código línea a línea y se actualiza el inventario de componentes del sistema, en segundo plano, sin parar el trabajo. La seguridad se desplaza al inicio del ciclo y deja de ser bloqueante.

IDENTIDAD Y ACCESOS

AHORA:

Hoy el uso de MFA no está ampliamente extendido, en ocasiones utiliza el SMS como canal de comunicación con el usuario, las credenciales son persistentes sin política de rotado y la revisión de permisos y accesos es un proceso manual.

A DONDE TENEMOS QUE IR:

La evolución del proceso debe incluir autenticación resistente a *phishing* (como FIDO2), accesos efímeros (como JIT) y un inventario vivo que también incluye las identidades de los agentes de IA.

COPIAS DE SEGURIDAD Y RECUPERACIÓN ANTE DESASTRES

AHORA:

Actualmente las políticas de copias de seguridad y recuperación ante desastres de los servicios críticos se prueban, en el mejor de los casos, una vez al año. Los planes y estrategias son documentos estáticos que no evolucionan o se actualizan salvo incidente grave que demuestre su falta de operatividad.

A DONDE TENEMOS QUE IR:

La capacidad de recuperación adquiere una importancia aún mayor. Las copias de seguridad deberán ser inmutables y estar aisladas de la red de producción. Las estrategias de resiliencia operativa deberán probarse con ejercicios reales, en periodos mucho más cortos de periodicidad.

RIESGO DE TERCEROS

AHORA:

Los proveedores se evalúan con cuestionarios periódicos y la evidencia que aportan no se actualiza hasta la siguiente revisión.

A DONDE TENEMOS QUE IR:

La **supervisión deberá ser continua**, apoyándose en **telemetría avanzada y artefactos de inventario como SBOM compartidos**, y evolucionado hacia modelos de **gestión continua del riesgo de terceros** que permitan un seguimiento dinámico de la exposición. Asimismo, resulta necesario incorporar prácticas específicas de **gestión del riesgo en la cadena de suministro de IA,** incluyendo el uso de **SBOM ampliados a sistemas de IA (*AI BOM)*** para mejorar la trazabilidad de modelos, dependencias y datos. En paralelo, los contratos con proveedores deben incorporar **cláusulas específicas sobre el uso de inteligencia artificial**, incluyendo evaluación de **modelos fundacionales**, gestión del riesgo asociado a proveedores de IA y obligaciones de transparencia, auditoría y notificación. Todo ello alineado con los **requisitos emergentes** del *AI Act* europeo, que refuerzan la necesidad de control sobre la cadena de valor de la IA, la gobernanza del dato y la supervisión continua de los sistemas desplegados.

4.2. Buenas prácticas para entornos OT

Debido a las particularidades del mundo de las Tecnologías de las Operaciones (OT), es conveniente tener en cuenta dichos entornos a la hora de valorar el impacto de la IA ofensiva, así como de definir las medidas de mitigación específicas de este ámbito.

- **Actualización del análisis de riesgos OT** incorporando explícitamente las capacidades de IA ofensiva como factor de amenaza. Los análisis existentes bajo IEC 62443-2-1 o NERC CIP deben revisarse para reflejar la reducción del umbral de conocimiento del atacante y la aceleración de fases de reconocimiento.

- **Definición de una política de gestión de superficie de ataque perimetral**: inventario continuo de dispositivos con exposición a Internet, criterios de riesgo aceptable, proceso de revisión periódica y responsabilidades claras de parcheo en dispositivos frontera IT-OT.

- **Evolución hacia un modelo de descubrimiento continuo y contextualizado de activos OT**, capaz de identificar automáticamente dispositivos, comunicaciones, dependencias operativas y criticidad de los procesos asociados, extendiendo la visibilidad a toda la cadena operacional

- Integración de la amenaza de IA ofensiva en los programas de formación y concienciación diferenciada para personal OT, **distinguiendo su perfil de riesgo específico del de un entorno IT corporativo.**

- **Potenciación de los laboratorios de acreditación para entornos OT**, considerando análisis específicos ante amenazas basadas en IA.

- **Revisión de los controles de seguridad en la cadena de suministro de *software* y *firmware* OT**, incluyendo mecanismos de verificación de integridad y procedimientos de actualización que reduzcan la ventana de exposición de dispositivos OT.

- **Segmentación reforzada entre IT y OT**, con particular atención a los equipos de electrónica de red en el interior de redes de control.

- **Despliegue de capacidad técnica de monitorización pasiva en redes OT**: sondas y puntos de captura que aporten visibilidad sobre el tráfico industrial

e inspección de los protocolos en uso en cada entorno, sin interferir con la disponibilidad del proceso. Esta capacidad es el sustrato técnico sobre el que se construyen las capacidades de detección.

Por último, una casuística especialmente relevante en entornos OT es la de los dispositivos que, por sus características técnicas o por restricciones operativas, **no admiten actualización ni parcheo**. En estos casos, la organización debe recurrir a controles compensatorios que reduzcan la exposición sin intervenir sobre el dispositivo.

[Fuente] S2 Grupo

4.3. Buenas prácticas por capacidades de ciberseguridad

El contexto actual de amenazas impulsadas por inteligencia artificial requiere evolucionar desde un enfoque tradicional de ciberseguridad basado en controles estáticos hacia un modelo estructurado por **capacidades operativas**, alineadas con la naturaleza dinámica y automatizada del ataque. En este sentido, la protección eficaz ya no depende únicamente de la implantación de medidas aisladas, sino de la capacidad de la organización para **detectar, prevenir, responder y adaptarse de forma continua** frente a un entorno de riesgo en constante evolución.

El presente apartado organiza las buenas prácticas en función de estas capacidades clave, proporcionando una visión integral y accionable de la seguridad, desde la identificación proactiva de vulnerabilidades hasta la respuesta automatizada ante incidentes. Este enfoque permite no solo reforzar la postura de seguridad, sino también facilitar la priorización de iniciativas, la asignación de recursos y la integración con marcos normativos como el Esquema Nacional de Seguridad (ENS).

De este modo, las capacidades se convierten en el eje vertebrador de la ciberseguridad, permitiendo a las organizaciones evolucionar hacia un modelo más resiliente, ágil y adaptado a la velocidad y sofisticación de las amenazas actuales.

Autoevaluación

La **autoevaluación** del nivel de protección frente a amenazas de IA ofensiva es un elemento crítico para que las organizaciones, especialmente en el sector público, comprendan su exposición real en un contexto donde la velocidad y la sofisticación del ataque superan los modelos tradicionales de defensa. Este ejercicio permite **identificar brechas en capacidades como detección, respuesta o gobierno de la IA, y priorizar medidas antes de que dichas debilidades sean explotadas**. Para llevarla a cabo de forma efectiva, las organizaciones deben apoyarse en procesos estructurados como evaluaciones de riesgo específicas de IA, revisiones continuas de la superficie de exposición, simulación de ataques sobre sistemas IA, así como en herramientas de monitorización, análisis de comportamiento e inteligencia y amenazas que integren señales en tiempo real. Además, pueden desarrollar marcos internos de evaluación basados en estándares, como ENS, ISO u OWASP, adaptados a IA, que permitan medir periódicamente su nivel de madurez y adecuar sus controles a la evolución constante de las amenazas.

Utilizar herramientas facilitadas por el **CCN-CERT** como el **formulario de autoevaluación disponible en INES** para conocer tanto el estado actual como las acciones recomendadas a ejecutar para mejorar el nivel de protección.

Cultura y capacitación

Reforzar el conocimiento organizativo sobre el impacto de la inteligencia artificial en la ciberseguridad, garantizando que los responsables comprenden tanto sus riesgos como sus implicaciones operativas. Forma a los equipos de ciberseguridad en los nuevos vectores de ataque, incluyendo la manipulación de instrucciones en modelos (*prompt-injection*), las suplantaciones mediante contenido manipulado (*deepfakes*), los agentes autónomos y la explotación automatizada. Capacitar tanto a perfiles técnicos como de gestión en el uso seguro de sistemas basados en IA, incorporando sus limitaciones y riesgos asociados. Integrar esta alfabetización en inteligencia artificial dentro de los programas de concienciación corporativa y extenderla al conjunto de los usuarios. Finalmente, desarrollar capacidades internas que permitan evaluar modelos, herramientas y agentes de IA desde una perspectiva de seguridad, asegurando un uso controlado, informado y alineado con el marco de riesgo de la organización.

- **Concienciación sobre ciberataques asistidos por IA**, suplantación hiperrealista y clonación de Voz.
- **Formación IA**: Desarrollo de materiales de formación de ciberseguridad IA.
- **Crear un AI *Red Team* interno y Operación de Vulnerabilidades** permanente como capacidades futuras.

Definir la estrategia y gobierno de la seguridad IA

Implantar un **modelo de gobernanza de la inteligencia artificial** que permita controlar de forma estructurada su uso, riesgos y capacidades dentro de la organización. Define políticas claras de utilización de la IA, estableciendo qué sistemas pueden emplearse, en qué contextos y bajo qué controles. Clasifica los sistemas según su nivel de criticidad y riesgo, aplicando medidas proporcionales en cada caso. Asigna responsabilidades específicas a lo largo de todo el ciclo de vida de los sistemas (desde el desarrollo hasta la operación y supervisión) y garantiza un control riguroso de accesos a modelos, agentes y datos, evitando usos indebidos o exposiciones no autorizadas. Integra esta gobernanza dentro del marco global de ciberseguridad, alineándola con los controles y procesos existentes, y establece mecanismos de supervisión continua y auditoría que permitan detectar desviaciones y asegurar un funcionamiento conforme a los requisitos definidos.

- Nivel de riesgo de IA aceptable y marco de referencia actualizado con 2ª línea.
- Crear un Comité IA o Foro de Innovación Segura. En el caso de que no sea una solución viable u operativa dentro de nuestra organización, integrar la gestión de la IA dentro de órganos ya existentes, como Comités de Seguridad de la Información, el Comités de Riesgos o Comités de Transformación Digital.
- CSA (*Cloud Security Alliance*) AI *Controls Matrix* integrada en proveedores e interna.

Proteger y reducir los tiempos de parcheo

Asumir que **la ventana entre la publicación de un parche y su explotación es mínima**, constituyendo un riesgo operativo inmediato que debe abordarse de forma prioritaria. Aplica el parcheo urgente de cualquier vulnerabilidad incluida en el catálogo de vulnerabilidades explotadas conocidas (KEV), tratando como incidente de seguridad aquellas con explotación activa expuestas a red. Complementa esta gestión con un enfoque de priorización basada en riesgo real, utilizando modelos como el Sistema de Puntuación para la Predicción de *Exploits* (EPSS) para ordenar el resto de las vulnerabilidades. Refuerza este enfoque con la implantación de autenticación multifactor resistente al *phishing* basada en FIDO2, eliminando métodos débiles como el SMS. Integra de forma transversal los principios de mínimo privilegio y confianza cero (*Zero Trust*), limitando accesos y verificando continuamente identidades y dispositivos, y adopta modelos como Plataforma de Protección de Aplicaciones Nativas en la Nube (CNAPP) para mejorar la visibilidad y protección en entornos *cloud*. Establece niveles de servicio exigentes, asegurando la aplicación de parches en menos de 24 horas en

sistemas expuestos a Internet y en plazos reducidos en el resto, y complementa este enfoque con capacidades de gestión continua de la postura de seguridad, tanto en IA (IA-SPM) como en aplicaciones SaaS (SSPM), para mantener un control permanente de la exposición. En este contexto, el parcheo deja de ser una opción negociable y pasa a ser una actividad crítica gestionada en función del impacto operativo de su despliegue.

- **MFA resistente a *Phishing***: FIDO2 es un estándar abierto de autenticación sin contraseña desarrollado por la FIDO *Alliance* y el W3C. Permite acceder a sitios web, aplicaciones y sistemas operativos utilizando criptografía asimétrica, biometría o llaves físicas. Es la mejor barrera contra el *phishing* y el robo de cuentas. Eliminar las credenciales estáticas.
- **Parcheo KEV en 24h**: corrección prioritaria de vulnerabilidades incluidas en el catálogo de **Explotaciones Conocidas y Activas (KEV)** de CISA. Estas vulnerabilidades se priorizan para el parcheo urgente porque ya están siendo explotadas por atacantes. Revisiones de código para identificar vulnerabilidades y SAST IA-asistido, que combina el escaneo estático de código tradicional con modelos de lenguaje para identificar y corregir vulnerabilidades. La IA actúa como asistente, priorizando alertas, reduciendo los falsos positivos y generando sugerencias de código para la remediación. Identificar los activos críticos para priorizarlos sobre el resto.
- Dimensionar el equipo o servicio de gestión de vulnerabilidades para escalar hasta un orden de magnitud superior.
- Incorporar los principios de mínimo privilegio, *Zero Trust* y CNAPP (protección de aplicaciones nativas en la nube).
- Utilizar marcos y herramientas para gestionar la postura de seguridad (AI-SPM, SSPM).

Detectar las vulnerabilidades antes de desplegar

La **detección** de vulnerabilidades antes del despliegue se ha convertido en un pilar esencial de la ciberseguridad en un entorno donde las amenazas evolucionan a gran velocidad y los atacantes emplean capacidades automatizadas basadas en inteligencia artificial. En este contexto, el tiempo disponible entre la identificación de una debilidad y su explotación efectiva se ha reducido drásticamente, pasando de días o semanas a horas o incluso minutos, lo que elimina prácticamente cualquier margen de actuación una vez los sistemas están en producción. Por ello, es imprescindible adoptar como premisa que toda vulnerabilidad que alcance entornos productivos será descubierta y potencialmente explotada por un adversario, lo que obliga a anticipar la defensa desde el origen.

Para responder a este escenario, las organizaciones deben **evolucionar hacia un enfoque en el que la seguridad forme parte intrínseca del ciclo de desarrollo**, integrándose de manera temprana y continua. Esto implica incorporar herramientas avanzadas de análisis de vulnerabilidades apoyadas en inteligencia artificial, capaces de examinar no solo el código fuente, sino también las configuraciones, dependencias y comportamientos de los sistemas. Estas herramientas permiten identificar debilidades que van más allá de los errores clásicos (como vulnerabilidades de lógica, malas configuraciones o cadenas de explotación complejas) que a menudo pasan desapercibidas para los mecanismos tradicionales.

Además del análisis, resulta clave reforzar la capacidad de reacción mediante la automatización de la corrección de vulnerabilidades. Esto se traduce en la generación automática de recomendaciones o incluso parches directamente aplicables a partir de los hallazgos detectados, especialmente aquellos procedentes de técnicas como el análisis estático, que permite revisar el código sin necesidad de ejecutarlo. En este proceso se debe incluir la revisión humana de posibles falsos positivos para evitar cambios que puedan suponer riesgos mayores que los que se busca mitigar. De esta forma, se acorta de manera significativa el tiempo entre la detección y su resolución, reduciendo el riesgo de que este pueda ser explotado.

Estas capacidades deben estar plenamente integradas en el ciclo de vida del desarrollo, formando parte de los procesos habituales de construcción, prueba y despliegue del *software*. Esto no solo permite detectar y corregir errores antes de que lleguen a producción, sino que también favorece la adopción de un modelo de seguridad preventiva y continua, en el que la identificación y mitigación de riesgos se realiza de forma constante y automatizada.

En conjunto, este enfoque permite reducir de manera significativa la **superficie de ataque**, limitar la exposición a amenazas y adaptar la seguridad al ritmo actual del ciberataque, caracterizado por su velocidad, automatización y capacidad de adaptación dinámica. En un entorno dominado por la inteligencia artificial, la anticipación deja de ser una ventaja para convertirse en una necesidad operativa imprescindible.

Revisar todo lo que ya se ha desplegado

La identificación y corrección de vulnerabilidades en código ya desplegado en producción constituye una actividad crítica en el contexto actual, donde los sistemas en funcionamiento representan el principal objetivo de los atacantes. A diferencia de las fases de desarrollo, el código en producción acumula dependencias, configuraciones y cambios históricos que incrementan la probabilidad de exposición a debilidades no detectadas previamente. Por ello, es imprescindible establecer procesos continuos de

revisión y análisis del *software* en operación, asumiendo que cualquier componente expuesto puede ser objeto de explotación.

En este proceso, resulta prioritario centrar los esfuerzos en aquellos componentes que presentan una mayor superficie de riesgo, como los **sistemas expuestos a Internet**, los módulos que gestionan entradas no confiables o las funcionalidades vinculadas a autenticación, autorización y auditoría, cuya explotación puede derivar en accesos indebidos o compromisos relevantes. Estos elementos deben ser objeto de revisiones específicas y recurrentes, ya que concentran gran parte del riesgo operativo y de impacto sobre la organización.

Asimismo, es especialmente relevante realizar auditorías sobre el **código heredado o sin mantenimiento activo**, que suele carecer de revisiones recientes y puede haber sido desarrollado bajo estándares de seguridad ya obsoletos. Este tipo de activos constituye, en muchos casos, uno de los principales vectores de entrada para atacantes, al combinar exposición con falta de vigilancia continua. La identificación de este código y su priorización en los planes de revisión es clave para reducir la exposición global.

Para que este enfoque sea efectivo, es necesario destinar una **capacidad de** ingeniería específica y sostenida para la **corrección de los hallazgos identificados**. No basta con detectar vulnerabilidades, la organización debe estar preparada para abordarlas con rapidez y eficacia, integrando estas tareas dentro de los ciclos habituales de mantenimiento y evolución del *software*. Esto implica asignar recursos, establecer prioridades claras y asegurar que las acciones correctivas se ejecutan en plazos adecuados, evitando la acumulación de deuda técnica.

En conjunto, este modelo permite avanzar hacia una **seguridad continua en producción**, en la que la revisión, identificación y corrección de vulnerabilidades forman parte del funcionamiento normal de los sistemas, reduciendo de forma progresiva la superficie de ataque y mejorando la resiliencia frente a amenazas cada vez más automatizadas y sofisticadas.

Diseñar asumiendo futuros compromisos y brechas de seguridad

Adoptar un **enfoque de diseño basado en la premisa de compromiso**, asumiendo que ningún sistema es completamente invulnerable y que, en algún momento, un atacante puede obtener acceso. Este cambio de paradigma obliga a priorizar la **contención del impacto y la resiliencia** por encima de la prevención absoluta, diseñando arquitecturas capaces de limitar la propagación de un incidente y reducir sus consecuencias operativas.

Para ello, es necesario implantar un modelo de **confianza cero *(Zero Trust)*** de forma efectiva, en el que cada interacción entre sistemas, servicios o usuarios sea tratada

como potencialmente no confiable. Esto implica autenticar y autorizar cada petición de manera explícita, independientemente de su origen, evitando confiar en perímetros tradicionales y asegurando la verificación continua de identidad y contexto. Este enfoque permite reducir significativamente el riesgo de movimientos laterales en caso de compromiso.

Complementar este modelo mediante la adopción de **mecanismos de autenticación robustos ligados al dispositivo**, de modo que el acceso no dependa únicamente de credenciales, sino también de la **confianza en el dispositivo utilizado**. Esto reduce drásticamente la exposición frente a ataques de suplantación y robo de credenciales.

Asimismo, sustituir el uso de **credenciales persistentes o secretos estáticos** por un modelo basado en **credenciales efímeras**, con validez limitada en el tiempo y generadas dinámicamente según la necesidad. La eliminación de claves de programación estáticas y otros secretos de larga duración es fundamental para evitar su reutilización por parte de un atacante en caso de fuga o compromiso.

Es importante **no limitar** la estrategia **a la aceleración de capacidades operativas**, como el escaneo o el parcheado, ya que la reducción de tiempos no resuelve los riesgos estructurales y puede comprometer la estabilidad de los sistemas si se omiten controles esenciales. Es imprescindible adoptar un **enfoque basado en la resiliencia**, orientado a **dificultar la explotación de vulnerabilidades incluso cuando estas existan**.

Normalizar conceptos básicos de diseño como el **aislamiento de componentes críticos, la gestión centralizada de correcciones** y la implantación de **defensas en profundidad**, diseñando un sistema donde la combinación de capas reduzca drásticamente la probabilidad y el impacto de un ataque, será más efectivo frente a las nuevas amenazas que añadir controles de forma indiscriminada.

En este sentido, aunque los modelos avanzados permiten automatizar y escalar la gestión de la ciberseguridad, no sustituyen el criterio experto ni la gobernanza, siendo clave evolucionar hacia sistemas seguros por defecto más que depender exclusivamente de la velocidad de respuesta.

En conjunto, este enfoque permite evolucionar hacia una arquitectura de seguridad en la que los sistemas están preparados para operar incluso bajo condiciones de compromiso, limitando la exposición, controlando los accesos y garantizando una mayor capacidad de respuesta ante amenazas avanzadas.

Gestionar en tiempo real la superficie de ataque

Asegurar la visibilidad y el control continuo de todos los activos expuestos, entendiendo que cualquier elemento no identificado o no gestionado representa un punto

potencial de entrada para el atacante. Para ello, establecer y mantener un inventario actualizado y dinámico de todos los hosts, servicios e interfaces de programación accesibles, incluyendo tanto infraestructuras tradicionales como entornos en la nube y componentes desplegados de forma descentralizada.

Complementar este control con una estrategia activa de **retirada de sistemas obsoletos o sin mantenimiento**, eliminando aquellos activos que carecen de responsable claro o que no pueden garantizar niveles adecuados de actualización y soporte. Este proceso debe ser deliberadamente exigente, ya que la permanencia de este tipo de sistemas incrementa significativamente la superficie de ataque y el riesgo operacional.

Adicionalmente, aplicar estrategias de reducción de la superficie de ataque, simplificando e implementando el principio de **exposición mínima necesaria**, asegurando que cada servicio o componente únicamente ofrezca las funcionalidades estrictamente imprescindibles para su operación. Esto implica revisar y limitar puertos abiertos, accesos externos, funcionalidades innecesarias y dependencias superfluas, reduciendo así la capacidad del atacante para identificar vectores explotables.

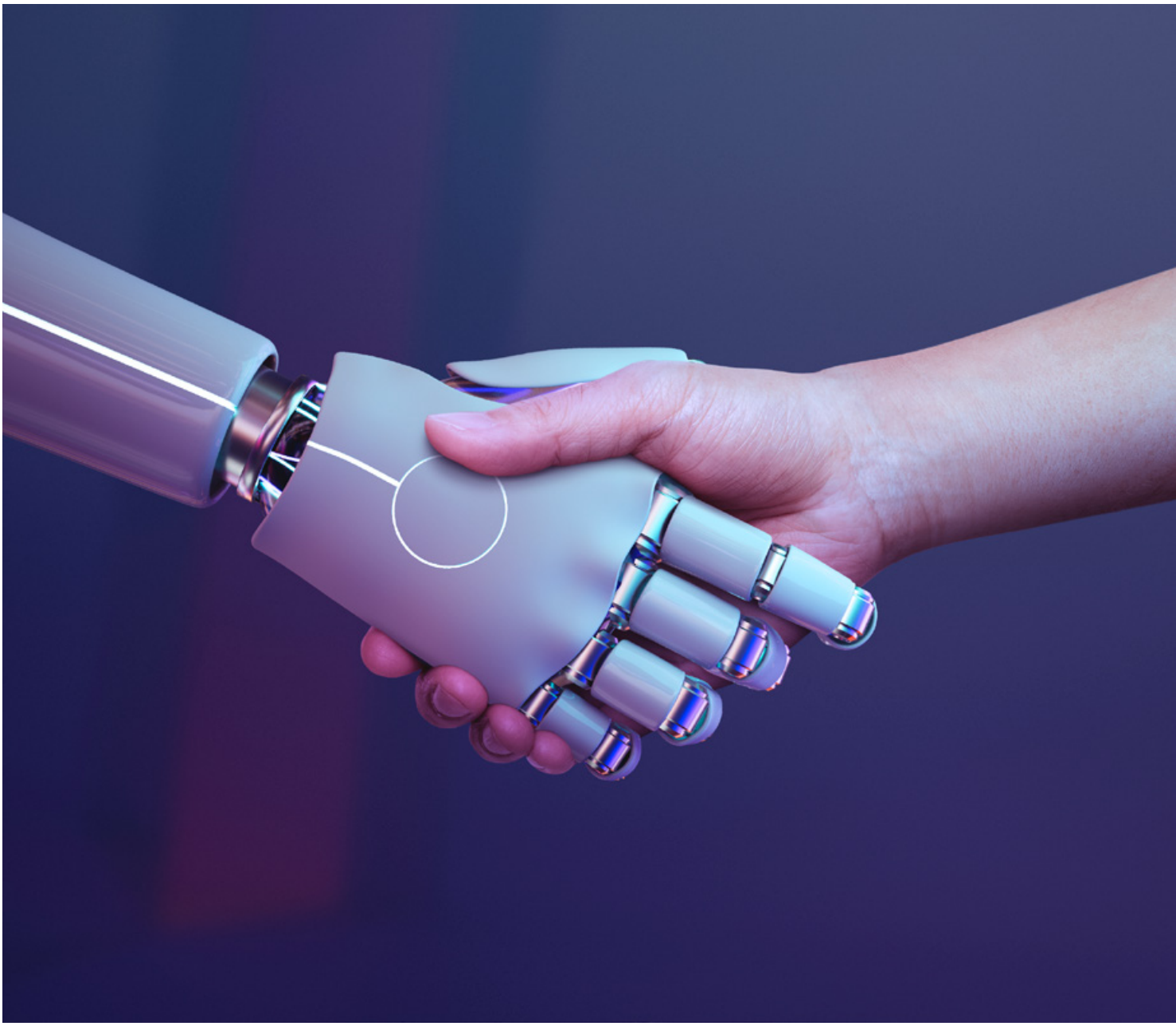
En conjunto, este enfoque permite evolucionar hacia una **gestión activa de la superficie de ataque**, en la que la organización mantiene un control permanente sobre sus activos, reduce la complejidad innecesaria y limita de forma efectiva las oportunidades de compromiso en un entorno cada vez más dinámico y distribuido.

Mejora de las capacidades de detección y respuesta

La reducción de los tiempos de respuesta ante incidentes se convierte en un factor crítico en el contexto de amenazas impulsadas por IA ofensiva, donde los ataques se desarrollan y escalan a velocidad máquina, reduciendo drásticamente el margen de reacción de las organizaciones. En este escenario, retrasos de minutos u horas pueden marcar la diferencia entre una contención eficaz y un compromiso completo del sistema, especialmente cuando los adversarios automatizan la explotación y el movimiento lateral. Para garantizar una respuesta eficaz, las organizaciones deben evolucionar hacia modelos de **detección y respuesta continua**, tanto en entornos **IT** como **OT**, apoyándose en herramientas basadas en IA agéntica capaces de correlacionar eventos en tiempo real, identificar comportamientos anómalos y ejecutar respuestas automatizadas, como SOAR, XDR o SOC autónomos. Estas capacidades permiten priorizar incidentes, reducir la intervención manual y activar contramedidas inmediatas como aislamiento de sistemas, revocación de credenciales o bloqueo de tráfico malicioso. Complementariamente, es esencial establecer procesos operativos como procedimientos automatizados, simulaciones de respuesta y ejercicios periódicos de resiliencia, integrando la inteligencia de amenazas en tiempo real para anticipar

vectores de ataque emergentes. En conjunto, la combinación de automatización, inteligencia contextual y orquestación permite reducir significativamente el tiempo de respuesta y adaptarlo a la velocidad actual del ciberataque.

En lo relativo a los entornos OT, dos tareas adicionales a las ya mencionadas adquieren especial relevancia: caracterización del comportamiento normal de los procesos mediante aprendizaje sobre el histórico operativo, y casos de uso de detección IT+OT verdaderamente correlacionados que reflejen las interdependencias ya existentes entre ambos entornos.



4.4. Seguridad del uso de agentes

Los agentes de inteligencia artificial representan una evolución significativa respecto a los sistemas tradicionales de IA, al incorporar capacidades de:

- Toma de decisiones autónoma
- Ejecución de acciones en sistemas
- Interacción continua con entornos digitales

Este nuevo paradigma introduce beneficios operativos relevantes, pero también amplía considerablemente la superficie de riesgo, al permitir que la IA **actúe directamente sobre sistemas críticos sin intervención humana directa**.

Los principales vectores de riesgo derivados del uso de agentes IA son:

- **Vectores de riesgo:**
  Los principales riesgos asociados al uso de agentes IA incluyen la ejecución no controlada de acciones, manipulación mediante *prompt-injection*, exfiltración de información, encadenamiento de decisiones erróneas, uso inseguro de APIs externas y falta de trazabilidad.

- **Principios de seguridad:**
  El uso seguro de agentes IA se basa en principios clave como el mínimo privilegio, la limitación de autonomía, la validación de entradas, la separación de capacidades, el control humano en decisiones críticas y la monitorización continua.

- **Controles técnicos:**
  Las medidas técnicas incluyen control de accesos, *sandboxing*, filtrado de entradas y salidas, gestión segura de credenciales, control de llamadas a APIs externas, registro y trazabilidad completa de las acciones del agente.

- **Relación entre riesgos, principios y controles:**
  (Ver la tabla de la siguiente página)

El uso de agentes IA requiere un enfoque específico de seguridad centrado en el control de acciones, la supervisión continua y la gobernanza. La adopción de estos principios y controles es esencial para garantizar su uso seguro en entornos críticos.

VECTOR DE RIESGO	PRINCIPIO DE SEGURIDAD	CONTROLES TÉCNICOS

Ejecución no controlada de acciones	Mínimo privilegio / limitación de autonomía.	Definir accesos estrictos y entornos aislados para limitar el impacto de los agentes IA:
		● Implementar Gestión de Identidades y Accesos (IAM) con roles mínimos (Azure AD, AWS IAM).
		● Uso de RBAC/ABAC para diferenciar permisos por tipo de agente.
		● Ejecutar agentes en entornos aislados:
		  • Contenedores *Docker* con permisos limitados
		  • Entornos en la nube *serverless* (AWS Lambda, *Azure Functions*)

Prompt injection	Validación de entradas / separación de capacidades.	Proteger al sistema frente a manipulación mediante entrada maliciosa:
		● Implementar Canalizaciones de validación de datos antes de enviar instrucciones (*prompts*) al modelo.
		● Uso de marcos de trabajo como:
		  • *Guardrails AI*
		  • *Microsoft Prompt Shields / Azure AI Content Filters*
		● Separar:
		  • Instrucciones del sistema
		  • Contexto del usuario
		● Aplicar sanitización de entradas (listas de control, detección de patrones maliciosos).

Exfiltración de datos	Protección de la información.	Evitar la fuga de información o respuestas no autorizada:
		● Implementar Prevención de Fugas de datos (*Data Loss Prevention* - DLP) en las respuestas.
		● Filtrado de contenido con:
		  • *Azure Content Safety*
		  • *OpenAI moderation endpoints*
		● Controlar acceso a datos mediante:
		  • APIs intermedias con validación
		  • Minimización de contexto (solo datos necesarios)
		● Uso de técnicas como reglas de validación de salidas (ej. no devolver credenciales, *tokens*, datos sensibles).

Errores encadenados	Control humano / limitación de autonomía.	Introducir supervisión humana en acciones críticas:
		● Flujo de aprobación obligatoria en:
		  • Cambios de configuración
		  • Accesos a datos sensibles
		● Integración con herramientas que permitan la supervisión/validación humana
		● Definir niveles de criticidad:
		  • Acciones automáticas (bajo riesgo)
		  • Acciones con aprobación (alto impacto)

Uso de APIs externas	Control de dependencias.	Controlar las interacciones externas del agente:
		● Uso de API *Gateway*.
		● Aplicar políticas.
		  • Lista de equipos permitidos
		  • Límite de respuestas
		  • Autenticación obligatoria
		● Monitorización para detección de uso anómalo de APIs.
		● Bloqueo de llamadas no aprobadas o fuera de patrón.

Gestión de credenciales	Protección de secretos.	Proteger secretos y evita exposición permanente:
		● Uso de sistemas de gestión de secretos.
		● Implementar:
		  • Credenciales efímeras (*Just-In-Time access*)
		  • Rotación automática de claves
		● Eliminar:
		  • API *keys* en el código
		  • Secretos en *prompts* o código

Falta de trazabilidad	Monitorización y auditoría.	Garantiza trazabilidad y capacidad de respuesta ante incidentes:
		● Registro detallado de:
		  • *Prompts* recibidos
		  • Decisiones del modelo
		  • Acciones ejecutadas
		● Integración con SIEMs.
		● Correlación de eventos para detectar:
		  • Comportamiento anómalo
		  • Posibles inyecciones de instrucciones maliciosas
		● Retención y auditoría para cumplimiento (ENS, ISO).

4.5. Integración de marcos de referencia en seguridad de IA

Para garantizar un enfoque sólido y alineado con las mejores prácticas internacionales, es necesario incorporar **marcos de referencia reconocidos** que permitan estructurar de forma coherente la seguridad en entornos basados en inteligencia artificial. Estos marcos aportan metodologías, controles y principios que facilitan tanto la implantación como la supervisión de medidas de seguridad adaptadas al nuevo contexto tecnológico.

En este sentido, resulta recomendable adoptar el ***"Multilayer Framework for Good Cybersecurity Practices for* AI" de ENISA**, que proporciona una estructura clara basada en tres niveles: fundamentos generales de ciberseguridad, controles específicos aplicados a sistemas de IA y medidas adaptadas a cada sector. Este enfoque facilita una visión integral que combina la protección tradicional con las necesidades particulares de la inteligencia artificial.

Asimismo, el **AI *Risk Management Framework* del NIST** constituye una referencia clave para el establecimiento de modelos de **gobernanza, gestión del riesgo y generación de confianza** en sistemas de IA. Este marco se ve reforzado por publicaciones más recientes del propio NIST, incluyendo perfiles específicos para IA generativa y análisis orientados a su uso en infraestructuras críticas, lo que amplía su aplicabilidad a entornos más complejos y sensibles.

En el ámbito de riesgos técnicos, resulta especialmente relevante incorporar las guías de OWASP para **aplicaciones basadas en modelos de lenguaje y sistemas de inteligencia artificial generativa**, que identifican vulnerabilidades comunes, como manipulación de entradas, fuga de información o abuso de capacidades de los sistemas. Estas referencias son fundamentales para abordar riesgos emergentes en aplicaciones basadas en IA y sistemas agénticos.

En el sector público español, el **Esquema Nacional de Seguridad constituye el marco normativo y operativo principal** para articular la protección frente a amenazas basadas o asistidas por inteligencia artificial. Las referencias internacionales —ENISA, NIST AI RMF, MITRE ATLAS, OWASP LLM/GenAI, ISO/IEC 42001 o *Cloud Security Alliance*— deben emplearse como instrumentos complementarios para enriquecer, actualizar y especializar los controles, pero no sustituyen la función vertebradora del ENS ni el papel del CCN como órgano técnico de referencia para su desarrollo, interpretación y aplicación práctica.

Esta integración alcanza también a las entidades privadas que utilicen sus propios sistemas de información y presten servicios o soluciones a entidades públicas en el ámbito de aplicación correspondiente, conforme al Esquema Nacional de Seguridad, aprobado por el Real Decreto 311/2022, de 3 de mayo.

En consecuencia, la seguridad frente a IA ofensiva debe integrarse en el análisis de riesgos ENS, en la categorización de sistemas, en la gestión de vulnerabilidades, en la protección de la información, en la trazabilidad, en la continuidad, en la gestión de proveedores, en la detección y respuesta ante incidentes y en la mejora continua. **La aparición de capacidades ofensivas basadas en IA no exige crear un marco paralelo, sino actualizar la aplicación del ENS** para que sus principios, requisitos mínimos y medidas de seguridad respondan a una amenaza más automatizada, adaptativa y rápida.

La **integración de estos marcos permite construir un modelo de seguridad robusto**, que combine el cumplimiento regulatorio con la adopción de buenas prácticas internacionales en un entorno marcado por la creciente influencia de la inteligencia artificial.

4.6. Hoja de ruta:

PRIMEROS PASOS

Cerrar la brecha de parcheo. Identidad resistente al *phishing*.

Sistemas críticos preparados: sin vulnerabilidades altas o críticas y en ciclos de evaluación continua.

Revisión de terceros: comenzar a evaluar a los terceros desde la perspectiva de la protección ante un escenario IA.

Segunda línea: planes de respuesta y marcos de referencia de riesgo actualizados. Estrategias de mitigación.

>>

PROPUESTA DE DISEÑO

Cambios de emergencia preautorizados.

Operaciones de vulnerabilidades como función permanente.

Detección a velocidad máquina.

Agilidad en la prueba y adquisición de elementos defensivos.

Diseño resiliente, aislado y seguro por defecto.

>>

CONSTRUCCIÓN DE LA CIBERSEGURIDAD AGÉNTICA

AI Red Team propio + externo.

Agentes defensivos gobernados.

Defender los agentes propios.

Concienciación específica sobre riesgos IA.

Marco de control IA.

>> PRIMEROS PASOS

- **Cerrar la brecha de parcheo:** Priorización del parcheo de vulnerabilidades sobre CISA KEV (Catálogo de vulnerabilidades explotadas conocidas) en 24 h, Utilizar la priorización según EPSS para el resto de las vulnerabilidades. Crear ventanas de cambio extraordinarias preautorizadas.
- **Identidad resistente al *phishing*:** La activación de autenticación multifactor debe cubrir la totalidad de los accesos remotos, sin excepciones. Donde no sea técnicamente viable de forma inmediata, la situación debe documentarse como riesgo aceptado con una medida compensatoria temporal y un plan de resolución fechado. Migrar a autenticaciones como FIDO2/*passkeys*, eliminar SMS-MFA, *tokens* efímeros. FIDO2 permite autenticarte de forma segura utilizando diferentes métodos:
  - **Dispositivo personal:** Usando tu propio ordenador o móvil mediante biometría (huella dactilar, reconocimiento facial) o un PIN local. A esto se le conoce comúnmente como *passkeys* (claves de acceso).
  - **Llaves físicas externas:** Dispositivos USB, NFC o *Bluetooth* (como *YubiKeys* o similares) que requieren estar físicamente presentes y, a menudo, el toque de un botón o un PIN para aprobar el inicio de sesión.
  - **Principales Ventajas:**
    - A prueba de *Phishing:* Los sitios web falsos no pueden engañar al sistema FIDO2. La llave o dispositivo solo responde si estás navegando en el dominio oficial.
    - Sin contraseñas maestras: Olvídate de recordar contraseñas largas o cambiarlas periódicamente.
    - Privacidad descentralizada: Los sitios web nunca almacenan un archivo de contraseñas que pueda ser *hackeado*.
    - *Zero Trust* en identidad y dispositivo.
- **Sistemas críticos preparados:** sin vulnerabilidades altas o críticas y en ciclos de evaluación continua. Se debe verificar la eliminación de credenciales por defecto en equipos de red, así como asegurar la existencia y accesibilidad de copias de configuración fuera de línea para los sistemas más críticos, con un procedimiento mínimo de recuperación documentado.
- **Revisión de terceros:** comenzar a evaluar a los terceros desde la perspectiva de la protección ante un escenario IA. Esta revisión debe incluir una auditoría de los accesos remotos de proveedores, verificando que operan bajo principios de mínimo privilegio y que las sesiones quedan auditadas.
- **Segunda línea:** planes de respuesta y marcos de riesgo actualizados. Definir estrategias de mitigación para las vulnerabilidades que no puedan ser parcheadas en los plazos adecuados.

"La IA no modifica nuestra estrategia ciber, pero sí acelera la necesidad de fortalecer lo básico."

>> REDISEÑOS QUE SE PROPONEN PARA ADAPTARSE A LAS AMENAZAS IA

- **Cambios de emergencia preautorizados:** Definir quién autoriza, en que casuísticas y en qué tiempos lo hace, algo a ensayar en ejercicios y pruebas de escenarios múltiples.
- **Gestión de Vulnerabilidades como función permanente:** descubrimiento continuo, parcheo asistido por IA, monitorización continua de la superficie de ataque.
- **Detección a velocidad máquina:** agentes para los casos de uso, contención preautorizada, Procedimientos de respuesta ante incidentes probados trimestralmente contra atacante con IA. Los procedimientos de respuesta deben incluir procedimientos específicos para los escenarios prioritarios del análisis de riesgos, y la capacidad de respuesta debe estar dimensionada para hacer frente a la velocidad de operación de la IA ofensiva, con mecanismos de contención que no requieran aprobación manual en los casos más críticos.
- **Agilidad en la prueba y adquisición de elementos defensivos:** Es necesario definir procesos ágiles para la incorporación de capacidades ciber. El factor velocidad introducido por la IA ofensiva hace que los ciclos tradicionales de evaluación y adquisición queden obsoletos; los procesos de validación y acreditación de nuevas capacidades defensivas deben rediseñarse para operar en plazos compatibles con el ritmo de evolución de la amenaza.
- **Implantación de defensas en profundidad por diseño y arquitectura:** Un enfoque de seguridad que consiste en no confiar en una única medida de protección, sino en establecer múltiples capas de controles complementarios que actúan de forma coordinada. El objetivo es que, si una de esas capas falla o es superada, existan otras que mitiguen el impacto, dificulten la progresión del ataque y protejan los activos críticos. Las vulnerabilidades y fallos son inevitables.

"Se diseñan entornos donde la explotación resulta compleja, limitada y detectable."

>> CONSTRUCCIÓN DE LA CIBERSEGURIDAD AGÉNTICA

- **AI *Red Team* propio + externo:** preparar los ejercicios de Pruebas de penetración basadas en amenazas (TLPT) con IA. Estos ejercicios deben incorporar simulaciones de campañas de ingeniería social adaptadas al contexto específico de cada organización, y sus resultados deben retroalimentar tanto el análisis de riesgos como el programa de formación.
- **Agentes defensivos gobernados:** Los agentes desplegados deben tener un alcance limitado, radio de impacto acotado, Interruptor de apagado, capacidad de supervisión e intervención humana, identidad definida y única por agente. El objetivo es disponer de una capacidad defensiva basada en IA que mejore de forma efectiva la prevención, detección y respuesta ante operaciones hostiles realizadas o apoyadas por IA, con un grado progresivo de autonomía gobernada.
- **Defender los agentes propios:** Hay que proteger los agentes IA adoptados de amenazas como peticiones u ordenes maliciosas (*prompt injection*), robo de credenciales del agente o manipulación de la herramienta, y se deben crear registros de auditoría.
- **Concienciación específica sobre riesgos IA:** La evolución de la inteligencia artificial ha multiplicado la eficacia de los ataques basados en ingeniería social avanzada, introduciendo formas de suplantación cada vez más difíciles de detectar. La suplantación hiperrealista mediante *deepfakes* permite recrear con precisión la imagen de directivos o empleados clave en entornos como videollamadas, con un alto nivel de credibilidad. La clonación de voz posibilita reproducir conversaciones completas con tono, ritmo y estilo realistas, facilitando la autorización fraudulenta de operaciones sensibles. Estas capacidades se combinan con el *phishing* asistido por IA, que genera mensajes altamente personalizados y adaptados al contexto del destinatario, incrementando de forma significativa su tasa de éxito. En paralelo, los atacantes pueden orquestar campañas dirigidas específicamente a perfiles con privilegios, como alta dirección o responsables financieros, utilizando información contextual y perfiles detallados para maximizar el impacto. Ante este escenario, resulta imprescindible reforzar los procesos de verificación mediante canales independientes (*out-of-band*), que permitan validar identidades y solicitudes críticas fuera del canal original, reduciendo el riesgo de suplantación y fraude. El programa de formación debe ser diferenciado por perfiles, con ejercicios prácticos que incluyan campañas de *phishing* y *vishing* simuladas con terminología y contexto propios de cada organización. La periodicidad debe ser suficiente para adaptarse a la evolución de las técnicas del adversario.
- **Ejemplos de marcos de control sobre la IA:** ISO/IEC 42001 establece un marco de gestión para sistemas de inteligencia artificial, orientado a garantizar su gobernanza, control de riesgos y cumplimiento regulatorio. El NIST AI RMF complementa este enfoque proporcionando una guía práctica para identificar, evaluar y mitigar riesgos asociados al ciclo de vida de la IA, con énfasis en fiabilidad, seguridad y ética. ATLAS (*Adversarial Threat Landscape for Artificial-Intelligence Systems*) aporta una taxonomía de amenazas específicas contra sistemas de IA, útil para anticipar vectores de ataque. OWASP LLM/ASI identifica vulnerabilidades y riesgos críticos en modelos de lenguaje y sistemas de IA generativa, facilitando su protección y, finalmente, el MRM (*Model Risk Management*) extendido a IA como amenaza amplía los modelos tradicionales de gestión de riesgo de modelos para incorporar tanto el riesgo operativo como el uso malicioso de la IA, integrando controles técnicos y de gobierno para mitigar impactos en negocio y cumplimiento.

Para ejecutar con garantías esta hoja de ruta, planteamos el siguiente plan dividido en 3 Fases a alto nivel:

FASE 1

Concienciación del riesgo y movilización de equipos:

En esta primera fase, es necesario impulsar una **toma de conciencia del riesgo asociado a la inteligencia artificial** y movilizar a la organización mediante la creación de un grupo de trabajo transversal que incluya a las áreas clave de **seguridad, tecnología, riesgos y cumplimiento.** Este equipo debe operar bajo un **modelo organizativo claro**, con mecanismos de **seguimiento continuo, espacios de coordinación intensiva y responsables definidos**. Sin un **liderazgo claro y un modelo de gobernanza activado desde el primer momento**, la organización responde a la IA ofensiva de forma reactiva y fragmentada, que es, precisamente, la postura que el adversario explota.

Paralelamente, se debe elaborar un **análisis de la situación actual** (estado del arte) para determinar el **nivel real de protección existente**, así como definir la **información y evidencias a solicitar a terceros** en materia de seguridad. Finalmente, resulta imprescindible formalizar un **registro específico de riesgos asociados a la inteligencia artificial** e integrar su gestión dentro de los **mecanismos de control y supervisión de la segunda línea de defensa**, asegurando su alineación con el modelo global de gestión de riesgos de la organización.

FASE 2

Construcción del programa y del plan de seguimiento:

En esta segunda fase, se aborda la **construcción del programa de seguridad** y la definición de un plan estructurado de seguimiento, adaptado a las necesidades y capacidades de la organización. Para ello, se debe desarrollar un **programa de actuación con plazos definidos y progresivos**, organizados en horizontes temporales claros (por ejemplo, de 0 a 45 días y de 45 a 90 días), que permitan una implantación ágil y controlada.

Asimismo, es necesario establecer una **oficina de seguimiento del programa,** encargada de coordinar las iniciativas y garantizar su avance mediante un **modelo de indicadores de control** que facilite la medición del progreso y la toma de decisiones. De forma complementaria, se deben habilitar **espacios de coordinación intensiva** específicos para aquellas iniciativas que impliquen cambios relevantes en los **modelos y procesos tradicionales de IT**, asegurando así una transición controlada y alineada con los objetivos estratégicos de la organización.

- Desarrollo de un **programa de seguridad adaptado a la entidad** con plazos acotados y reducidos:
  - **Contener:** Prioriza el parcheo inmediato de vulnerabilidades críticas, impide despliegues con fallos relevantes y refuerza el desarrollo con detección automatizada mediante IA. Completa este enfoque con alineación con proveedores y planes de contingencia adaptados a ataques con IA, asegurando preparación operativa frente a nuevas amenazas.
  - **Construir:** Implanta gestión continua de exposición, integra seguridad en el desarrollo con IA y refuerza los controles básicos (confianza cero, mínimo privilegio y segmentación). Complementa con gestión de terceros actualizada, detección avanzada apoyada en inteligencia de amenazas y protección de copias de seguridad resistentes a ataques.
  - **Madurar:** Implanta pruebas continuas de seguridad basadas en IA, refuerza el gobierno del dato en los flujos de IA y evoluciona los controles hacia arquitecturas avanzadas (confianza cero, detección de amenazas de identidad y aislamiento de navegación). Incorpora capacidades de operación autónoma en el centro de seguridad, análisis predictivo de ataques y respuesta automatizada a velocidad máquina, apoyadas en procedimientos adaptados a amenazas con IA. Todo ello bajo marcos de referencia específicos y con el impulso de roles especializados que lideren la adopción de la ciberseguridad basada en IA.
  - **Horizonte:** Implanta una operación continua de gestión de vulnerabilidades y pruebas internas con IA, elimina el riesgo

arquitecturas capaces de anticipar y autorrecuperarse. Incorpora modelos avanzados como gemelos digitales y operaciones agénticas seguras, con identidades dinámicas y acceso bajo demanda. Refuerza la cadena de suministro con capacidades tempranas de alerta y habilita recuperación automática asistida por IA, alineada con requisitos de resiliencia operativa.

- Creación de una **oficina de seguimiento del programa** y desarrollo de un modelo de indicadores de seguimiento.
- **Grupo de trabajo** específicas para las iniciativas que afecten a los modelos y procedimientos tradicionales de IT.

Implanta pruebas continuas de seguridad basadas en IA, refuerza el gobierno del dato en los flujos de IA y evoluciona los controles hacia arquitecturas avanzadas (confianza cero, detección de amenazas de identidad y aislamiento de navegación). Incorpora capacidades de operación autónoma en el centro de seguridad, análisis predictivo de ataques y respuesta automatizada a velocidad máquina, apoyadas en procedimientos adaptados a amenazas con IA. Todo ello bajo marcos de referencia específicos y con el impulso de roles especializados que lideren la adopción de la ciberseguridad basada en IA.

FASE 3

Concienciación del riesgo y movilización de equipos:

En esta tercera fase se aborda la **alineación, validación y puesta en marcha del programa** con los principales grupos de interés de la organización. Para ello, es necesario presentar un resumen ejecutivo del programa y su evolución, dirigido a la alta dirección y al consejo, asegurando su comprensión y apoyo. Asimismo, se debe detallar de forma clara el **impacto esperado sobre los procesos de IT**, especialmente en aquellos ámbitos que requieren transformación o adaptación.

De forma complementaria, resulta fundamental preparar un **plan de respuesta coordinado** ante posibles requerimientos de **reguladores y clientes**, garantizando coherencia y transparencia. Además, se debe evaluar y, en su caso, establecer un **comité específico de riesgos de inteligencia artificial,** o integrar esta materia en el **comité de riesgos existente**, asegurando un gobierno adecuado. Finalmente, esta fase culmina con la aprobación formal del programa y el inicio de su ejecución operativa, consolidando su despliegue en la organización.

4.7. Decálogo de recomendaciones

1 Autoevaluar el nivel de protección

Definir un modelo de evaluación continuo del riesgo específico de IA, integrando capacidades de detección, respuesta y gobernanza. Con el formulario disponible en INES puedes evaluar tu nivel de riesgo actual específico frente a la IA.

2 Crear cultura en IA

Fortalecer el conocimiento organizativo en torno al impacto de la IA en la ciberseguridad, asegurando que los responsables comprenden tanto sus riesgos como sus implicaciones operativas.

3 Establecer la gobernanza de la IA como pilar básico

Implantar un modelo de gobernanza que permita controlar el uso, los riesgos y las capacidades de la IA dentro de la organización de forma estructurada.

4 Reducir el gap en el parcheo

Asumir que la ventana entre publicación del parche y explotación es mínima y constituye un riesgo operativo real. Prioriza los activos críticos identificados y las vulnerabilidades potencialmente explotables.

5 Prepararse para un aumento masivo de vulnerabilidades reportadas

Dimensionar el equipo o servicio de gestión de vulnerabilidades para escalar hasta un orden de magnitud superior.

6 Detectar vulnerabilidades antes de desplegar

Asumir que cualquier vulnerabilidad en producción será descubierta antes por un atacante.

7 Revisar continuamente el código ya desplegado

Identificar y corregir vulnerabilidades existentes en código que ya se encuentra en producción.

8 Diseñar los sistemas asumiendo compromiso

Adoptar un enfoque de diseño basado en la premisa de que el sistema será comprometido.

9 Gestionar activamente la superficie de ataque

Asegurar la visibilidad y control continuo de todos los activos expuestos. Los organismos de la Administración pueden hacerlo con ELSA.

10 Reducir los tiempos de respuesta ante incidentes

Adaptar las capacidades de respuesta al ritmo de las amenazas actuales. Con la ventaja competitiva que supone la IA desde el punto de vista ofensivo, la capacidad de recuperación y resiliencia es más crítica que nunca.

5. ANEXO A. Glosario

TÉRMINO / ACRÓNIMO	SIGNIFICADO
AI Red Team	Equipo o capacidad de pruebas ofensivas especializada en sistemas de IA, modelos, agentes y su ecosistema.
AI RMF	*AI Risk Management Framework.* Marco de gestión de riesgos de IA de NIST.
AI-SPM	*AI Security Posture Management.* Gestión de la postura de seguridad de sistemas y servicios de IA.
API	*Application Programming Interface.* Interfaz de programación de aplicaciones.
ATLAS	*Adversarial Threat Landscape for Artificial-Intelligence Systems.* Base de conocimiento de MITRE sobre tácticas y técnicas adversarias contra sistemas de IA.
CISA	*Cybersecurity and Infrastructure Security Agency.* Agencia estadounidense de ciberseguridad y seguridad de infraestructuras.
CNAP	*Cloud-Native Application Protection Platform.* Plataforma integrada de protección para aplicaciones cloud-native.
CSA	*Cloud Security Alliance.* Organización de referencia en seguridad *cloud.*
CVE	*Common Vulnerabilities and Exposures.* Identificador normalizado de vulnerabilidades conocidas.
CVSS	*Common Vulnerability Scoring System.* Sistema de puntuación de severidad de vulnerabilidades.
Deepfake	Contenido sintético generado o manipulado con IA para imitar voz, imagen o vídeo de una persona real.
ENS	Esquema Nacional de Seguridad.

5. ANEXO A. Glosario

TÉRMINO / ACRÓNIMO	SIGNIFICADO
EPSS	*Exploit Prediction Scoring System.* Sistema predictivo que estima la probabilidad de explotación de una vulnerabilidad en los próximos 30 días.
FIDO2	*Fast Identity Online 2.* Estándar de autenticación fuerte y resistente al *phishing.*
IA	Inteligencia artificial.
IA agéntica	Sistema de IA capaz de percibir su entorno, tomar decisiones y ejecutar acciones de forma autónoma para alcanzar objetivos, con mínima intervención humana.
IA generativa	Sistema de IA capaz de producir contenido nuevo —texto, imagen, audio o código— a partir de patrones aprendidos durante su entrenamiento.
IA ofensiva	Uso de sistemas de IA para planificar, acelerar o ejecutar ciberataques, incrementando su velocidad, escala y sofisticación y reduciendo la barrera de entrada del atacante.
ISO/IEC 42001	Norma internacional para sistemas de gestión de inteligencia artificial.
IT	*Information Technology.* Tecnologías de la información.
JIT	*Just-In-Time.* Concesión de acceso o privilegios solo cuando se necesitan y por tiempo limitado.
KEV	*Known Exploited Vulnerabilities.* Catálogo de vulnerabilidades conocidas explotadas activamente mantenido por CISA.
LLM	*Large Language Model.* Modelo de lenguaje de gran tamaño.
Logging	Registro de eventos, acciones o transacciones para auditoría, monitorización e investigación.
MFA	*Multi-Factor Authentication.* Autenticación multifactor.
MRM	*Model Risk Management.* Gestión del riesgo de modelos.
NIST	*National Institute of Standards and Technology.* Organismo estadounidense de estandarización y referencia técnica.
OWASP	*Open Worldwide Application Security Project.* Comunidad y referencia global en seguridad de aplicaciones.
OWASP ASI	*OWASP Agentic Security Initiative.* Iniciativa de OWASP centrada en los riesgos y controles de seguridad para aplicaciones y agentes autónomos.
Passkeys	Credenciales modernas basadas en criptografía asimétrica que sustituyen contraseñas tradicionales.

5. ANEXO A. Glosario

TÉRMINO / ACRÓNIMO	SIGNIFICADO
Pipeline	Cadena automatizada de procesos de desarrollo, validación, pruebas y despliegue.
Prompt injection	Ataque que manipula instrucciones o contexto para alterar el comportamiento previsto de un modelo o agente.
Sandboxing	Ejecución aislada de procesos o código en un entorno controlado para limitar impacto.
SAST	*Static Application Security Testing.* Análisis estático de seguridad del código fuente.
SBOM	*Software Bill of Materials.* Inventario estructurado de componentes software de una aplicación.
Shadow AI	Uso de herramientas o servicios de IA fuera del gobierno, inventario o control oficial de la organización.
SIEM	*Security Information and Event Management.* Plataforma de gestión y correlación de eventos e información de seguridad.
SMS - MFA	*Autenticación multifactor* basada en mensajes SMS.
SOC	*Security Operations Center.* Centro de operaciones de seguridad.
SOAR	*Security Orchestration, Automation and Response.* Orquestación, automatización y respuesta de seguridad.
SSPM	*SaaS Security Posture Management.* Gestión de la postura de seguridad en aplicaciones SaaS.
Threat intelligence	Inteligencia de amenazas. Información analizada sobre actores, campañas, técnicas, vulnerabilidades e indicadores de compromiso.
TLPT	*Threat-Led Penetration Testing.* Prueba avanzada de intrusión guiada por inteligencia de amenazas y basada en tácticas realistas de adversarios.
War Room	Espacio o dinámica de coordinación intensiva para gestión de crisis, incidentes o iniciativas críticas.
XDR	*Extended Detection and Response.* Detección y respuesta extendidas a múltiples capas de seguridad.
***Zero Trust* / ZT**	Modelo de seguridad que no confía por defecto en ningún usuario, dispositivo o flujo, y exige verificación continua.


JUNIO 2026
CCN-CERT
BP/36

centro criptológico nacional

Guía de buenas prácticas frente al modelo de IA ofensiva

GUÍA DE SEGURIDAD

www.ccn.cni.es

www.ccn-cert.cni.es

oc.ccn.cni.es