Medidas de **ciberseguridad y ciberresiliencia** ante modelos de **IA de frontera**

Información para entidades

incibe-cert_

GOBIERNO DE ESPAÑA
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES

incibe_
INSTITUTO NACIONAL DE CIBERSEGURIDAD

17 de junio de 2026

Medidas_ciberseguridad_ciberresiliencia_modelos_frontera.pdf

ÍNDICE

1. CONTEXTO ACTUAL

Los grandes modelos de lenguaje (*Large Language Model*, LLM), también conocidos como “modelos de IA” (*AI models*), son sistemas de inteligencia artificial entrenados con enormes volúmenes de datos y capaces de realizar tareas complejas como razonamiento, generación de código, análisis multimodal o automatización de procesos. Múltiples empresas tecnológicas compiten activamente en el desarrollo de modelos cada vez más potentes y accesibles.

En los últimos meses, han ganado relevancia en el ámbito de la ciberseguridad los denominados modelos de IA de frontera (*frontier AI models).* Cabe destacar que la utilización de LLM avanzados supone una ventaja tanto para defensores como para atacantes. Por un lado, se puede utilizar IA avanzada para detectar amenazas en tiempo real, automatizar respuestas ante incidentes y analizar grandes cantidades de eventos de seguridad a una escala y velocidad superiores a las del análisis manual, ayudando a reducir tiempos de respuesta y a mejorar las capacidades operativas de los centros de seguridad.

Sin embargo, estas mismas capacidades también están siendo aprovechadas por actores maliciosos. Los ciberdelincuentes pueden utilizar modelos avanzados para generar campañas de *phishing* más convincentes, automatizar la creación de *malware* o explotar vulnerabilidades de los sistemas. Esto reduce la barrera técnica para ejecutar ataques complejos y acelera la sofisticación del cibercrimen, obligando a empresas y gobiernos a replantear sus estrategias defensivas.

La mayor diferencia que supone es que el acortamiento de los tiempos en la detección de vulnerabilidades exige a su vez una reducción en los procesos de actualización y parcheo de los sistemas, que no siempre están preparados. Por otro lado, las capacidades de los nuevos modelos de encadenar distintas vulnerabilidades hacen que ya no se pueda relegar de la misma forma la corrección de una vulnerabilidad en base a su criticidad evaluada de forma independiente, sino que deba considerarse también el impacto que podría llegar a tener combinada con otras vulnerabilidades (conocidas o potenciales). Esto se refleja principalmente en una mayor presión en los equipos de seguridad, operación, desarrollo y respuesta, que muchas veces no se encuentran adecuadamente dimensionados a esta nueva realidad.

2. MEDIDAS DE CIBERSEGURIDAD Y CIBERRESILIENCIA

Desde INCIBE-CERT le remitimos medidas de ciberseguridad y ciberresiliencia que pueden ayudarle a proteger su organización ante incidentes y ciberataques que puedan derivarse del uso de modelos de inteligencia artificial de frontera y poner en riesgo su información y la continuidad de sus servicios y actividades.

En cualquier caso, dichas medidas y recomendaciones no sustituyen a las buenas prácticas que se vienen trasladando desde esta entidad, como autenticación de usuarios internos y externos, seguridad del software, defensa perimetral, sistemas de respaldo, etc., sino que son un refuerzo y/o complemento de las mismas.

Las medidas se agrupan en dos planos complementarios. El primero refuerza la protección de la organización frente a actividades ofensivas asistidas por inteligencia artificial. El segundo regula el uso seguro de la propia inteligencia artificial dentro de la organización, tanto para fines defensivos como en su actividad ordinaria. Las capacidades técnicas son con frecuencia las mismas para una y otra finalidad, por lo que su gobernanza resulta determinante.

Gestión de identidades y accesos frente a fuerza bruta e ingeniería social asistida por IA

La IA permite a los atacantes crear campañas de *phishing* hiperpersonalizadas, clonar voces/video (*deepfakes*) y automatizar ataques de diccionario adaptativos. La defensa debe asumir un aumento de la plausibilidad y la frecuencia de los intentos de suplantación.

- **Autenticación Adaptativa y MFA Robusto**. Implementar autenticación multifactor (MFA/2FA) obligatoria en todos los accesos (especialmente remotos y de terceros). Se deben priorizar métodos resistentes al *phishing* (como llaves físicas o FIDO2) frente a los SMS o códigos tradicionales, los cuales pueden ser interceptados o evadidos mediante MITM automatizados.
- **Arquitectura de Confianza Cero (*Zero Trust*)**. Eliminar la confianza implícita dentro de la red. Cada solicitud de acceso debe ser verificada continuamente en función del contexto (dispositivo, ubicación, comportamiento), limitando los movimientos de posibles agentes de IA autónomos que logren vulnerar el perímetro.
- **Mitigación de fuerza bruta avanzada**. Reemplazar sistemas de autenticación antiguos basados en el almacenamiento de usuario-contraseña por flujos modernos como OAuth o el uso de *tokens* específicos por usuario y servicio. Es crítico incorporar mecanismos dinámicos de limitación de peticiones (*throttling*) y alertas automáticas, ya que la IA puede modular los ataques de fuerza bruta para simular un comportamiento humano y evitar límites estáticos.
- **Control de privilegios riguroso (PAM/RBAC)**. Restringir los accesos bajo el principio de mínimo privilegio mediante soluciones de gestión de cuentas privilegiadas (PAM). Si un *infostealer* compromete credenciales con permisos de administrador, un atacante asistido por IA puede acelerar el reconocimiento interno, la escalada de privilegios y el movimiento lateral.
- **Capacitación contra *phishing* de nueva generación**. Actualizar los programas de concienciación de los empleados. Las simulaciones tradicionales basadas en "*faltas de ortografía*" han perdido mucha utilidad; se debe entrenar al personal para

identificar pretextos sofisticados generados por LLM y establecer protocolos de verificación ante posibles clonaciones de voz.

Seguridad del software y gestión del *Time-to-Exploit*

Las herramientas de IA ofensivas pueden acortar el tiempo entre la publicación de una vulnerabilidad y la disponibilidad de un *exploit* funcional, según el caso. Ello reduce el margen para mantener ventanas prolongadas de exposición en activos críticos o expuestos a Internet.

- **Priorización contextual de vulnerabilidades**. No limitar la priorización a la severidad CVSS individual. Incorporar criterios de exposición a Internet, criticidad del activo, existencia de explotación activa o de prueba de concepto, dependencia entre sistemas y posibilidad de encadenar varias vulnerabilidades para el movimiento lateral.
- **Reducción de ventanas de exposición**. Mantener una política continua de actualizaciones, reduciendo el tiempo de mitigación en activos expuestos a Internet a un máximo de 24-48 horas para vulnerabilidades críticas. Si el parche no puede aplicarse inmediatamente, se deben usar reglas automáticas de mitigación en los sistemas perimetrales de seguridad como WAF, firewalls o EDR, o deshabilitar el acceso al servicio hasta que pueda ser remediado.
- **Auditoría continua y reducción de la superficie**. Mantener inventarios precisos y automatizados de activos. Desmantelar sistemas heredados, cerrar puertos no utilizados y deshabilitar servicios innecesarios para evitar que herramientas automatizadas de reconocimiento identifiquen servicios expuestos, configuraciones débiles o activos no inventariados.
- **Seguridad en el desarrollo** (DevSecOps + IA controlada). Integrar herramientas automatizadas en los flujos de integración y despliegue continuo (CI/CD): análisis de composición de software (SCA) para vulnerabilidades en dependencias de terceros; análisis estático y dinámico (SAST/DAST) sobre el código propio; y revisión manual para los fallos de lógica de autorización (como IDOR o BOLA), que las herramientas automáticas no detectan de forma fiable.
- **Gobernanza sobre asistentes de código** (*copilots*). El uso de IA para escribir *software* corporativo debe estar estrictamente supervisado. Es obligatorio validar técnicamente cualquier código generado por IA, ya que estos modelos pueden introducir vulnerabilidades, fallas de funcionalidad, o sugerir dependencias inexistentes (*AI package hallucination*).

Uso de inteligencia artificial para la detección y corrección de vulnerabilidades

La inteligencia artificial no solo refuerza las capacidades del atacante; también está al alcance del defensor para identificar y corregir vulnerabilidades en su propio software. Conviene subrayar que este uso defensivo ya no depende del acceso a modelos avanzados de difícil obtención: existen capacidades de uso generalizado, integrables en el ciclo ordinario de desarrollo y mantenimiento, accesibles para la mayoría de las organizaciones.

- **Análisis asistido de código y de dependencias.** Incorporar herramientas que analizan el código y sus componentes en busca de patrones vulnerables, dependencias desactualizadas o con vulnerabilidades conocidas, y configuraciones inseguras. El análisis de composición de software (SCA), el escaneo de artefactos e

imágenes y las pruebas automatizadas de *fuzzing* permiten detectar de forma continua tanto vulnerabilidades conocidas como defectos no documentados, e integrarse como puntos de control previos al despliegue.

- **Asistentes de desarrollo supervisados.** Los asistentes basados en IA pueden explicar hallazgos, sugerir correcciones y ayudar a priorizar el trabajo de remediación. Deben emplearse como apoyo, nunca como auditoría completa: su salida requiere validación técnica, dado que pueden pasar por alto fallos de lógica de autorización o introducir errores. La supervisión humana sobre el código generado o corregido por IA es obligatoria.
- **Modalidades de despliegue.** Estas capacidades se ofrecen en formatos diversos, como servicios en la nube pública, plataformas integradas de desarrollo, asistentes conversacionales y soluciones desplegables en las propias instalaciones (*on premise*) o en entornos de nube soberana o controlada. La elección debe atender a la confidencialidad del código y de los datos, a los requisitos de residencia y jurisdicción, y al grado de integración con la plataforma de desarrollo de la organización, que puede ser un factor determinante de eficacia. Estas mismas exigencias deben trasladarse contractualmente al desarrollo y mantenimiento de software subcontratado.
- **Naturaleza dual de la capacidad.** Las herramientas que permiten al defensor analizar código, dependencias y artefactos son, en buena medida, las mismas que reducen la barrera para el atacante en el reconocimiento, el análisis de código expuesto o filtrado y la identificación de patrones explotables. Esta dualidad refuerza la necesidad de integrar estas capacidades en el ciclo de vida del software, gobernar su uso y reducir las ventanas de exposición, en lugar de confiar en su mera disponibilidad.

Defensa perimetral y monitorización (frente a la evasión de firmas)

Las amenazas asistidas por IA pueden modificar dinámicamente la estructura de sus cargas útiles (*malware* polimórfico) para evadir las firmas de los antivirus tradicionales.

- **Detección basada en comportamiento y anomalías**. Desplegar soluciones de seguridad perimetral (WAF avanzado, API Gateway) y de *endpoint* (EDR/XDR) que no dependan únicamente de firmas estáticas, sino de análisis de comportamiento y detección de anomalías en tiempo real.
- **Límites de abuso en APIs**. Definir umbrales estrictos de peticiones por dirección IP, usuario y clave de API para detener las actividades de enumeración masiva y *scraping* automatizado ejecutados por *bots* de IA.
- **Segmentación y microsegmentación de red**. Aislar las cargas de trabajo críticas. Esto actúa como un cortafuegos interno que impide que una intrusión automatizada se propague lateralmente de forma rápida dentro de la organización.
- **Threat Hunting activo**. Implementar capacidades de búsqueda proactiva de amenazas para identificar indicadores tempranos de compromiso antes de que los ataques se consoliden.

Inteligencia de amenazas y respuesta dinámica

La velocidad de los ataques modernos requiere que las organizaciones consuman y compartan datos de amenazas de manera automatizada.

- **Ingesta automatizada de IoC**. Los Indicadores de Compromiso (direcciones IP maliciosas, *hashes* y otras reglas de detección de *malware*, etc.) deben ser consumidos e integrados automáticamente mediante API en los sistemas de defensa.
- **Colaboración en ecosistemas de confianza**. Fomentar la compartición bidireccional de los IoCs detectados internamente en redes sectoriales (ISAC)[1] y con organismos de referencia como INCIBE-CERT (plataforma ICARO[2]) para inmunizar al tejido empresarial de forma colectiva.
- **Orquestación de la respuesta (SOAR)**. Adaptar los planes de respuesta a incidentes para escenarios de alta velocidad. Se deben delegar ciertas acciones de contención (aislar un servidor, revocar un *token*) a sistemas automatizados, ya que los tiempos de reacción humanos pueden resultar insuficientes ante ataques asistidos o acelerados mediante automatización.

Resiliencia operativa y cumplimiento regulatorio

- **Estrategia de respaldos a prueba de *ransomware***. Las copias de seguridad (locales y remotas) deben revisarse periódicamente y seguir la regla de inmutabilidad. Las herramientas empleadas por los atacantes pueden buscar activamente los servidores de copia de seguridad para inutilizarlos antes de cifrar los datos de la organización.
- **Gobernanza y cumplimiento legal**. Alinear la estrategia tecnológica con marcos internacionales, como el *Cybersecurity Framework (CSF) 2.0* y *AI Risk Management Framework*[3] de NIST, o *Multilayer Framework for Good Cybersecurity Practices for AI* [4] de ENISA. Las empresas deben prever el cumplimiento de las normativas vigentes, como la Ley de IA de la Unión Europea[5] y el Reglamento de Ciberresiliencia (CRA),[6] garantizando la transparencia y la seguridad desde el diseño. La aplicabilidad concreta de estas normas dependerá del papel de la entidad (usuaria, desarrolladora, fabricante, proveedora u operadora) y del tipo de sistema, producto o servicio afectado.

Uso seguro de la inteligencia artificial en la organización

- **Seguridad de la información.** Es fundamental tener en cuenta la confidencialidad de la información que se comparte con estos modelos y si es acorde con la política de riesgos establecida por la entidad, en función del tipo de modelo, proveedor que procesa los datos, acuerdos firmados, legislación aplicable, etc.
- **Riesgo de *prompt injection.*** Los modelos de IA pueden ser influenciados por instrucciones maliciosas o no autorizadas incluidas en los datos de entrada (documentos, correos electrónicos, páginas web, etc.), por lo que cuando se integra

[1] https://www.incibe.es/incibe-cert/sectores-estrategicos/ES-ISAC

[2] https://www.incibe.es/incibe-cert/servicios-operadores/icaro

[3] https://www.nist.gov/itl/ai-risk-management-framework

[4] https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai

[5] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[6] https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=OJ:L_202402847

IA en los procesos empresariales, es necesario tener en cuenta que estas instrucciones pueden alterar el comportamiento esperado del sistema.

Finalmente, se recomienda a las organizaciones mantener un estado de vigilancia reforzada y revisar de forma periódica la correcta aplicación de estas medidas, integrándolas dentro de sus procesos habituales de gestión de la ciberseguridad y continuidad de negocio. La adopción de controles preventivos, la monitorización activa y la colaboración con organismos especializados como INCIBE-CERT contribuyen de forma significativa a mejorar la capacidad de detección y respuesta ante posibles amenazas.