RETECH
Redes Territoriales de Especialización Tecnológica

RED ARGOS

Guía de respuesta a incidentes de ciberseguridad en sistemas IA en entornos de Salud y Smart Cities

Febrero de 2026

Versión 1.0

Junta de Andalucía

Financiado por la Unión Europea
NextGenerationEU

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

Plan de Recuperación, Transformación y Resiliencia

incibe_
INSTITUTO NACIONAL DE CIBERSEGURIDAD

Junta de Andalucía

Agencia Digital de Andalucía

Objeto del Expediente:

“ELABORACIÓN DE GUÍAS DE CIBERSEGURIDAD IA PARA ENTORNOS DE SMART CITIES Y SALUD” (CONTR 2024 829535).

Esta guía forma parte de la colección Guías de Ciberseguridad en IA para las Smart Cities y el sector Salud creada por la Agencia Digital de Andalucía como parte del Proyecto Red Argos, perteneciente al programa RETECH, de Redes Inteligentes de Especialización Tecnológica, en el marco del Plan de Recuperación, Transformación y Resiliencia, financiado por la Unión Europea-NextGenerationEU, a través de INCIBE.

Autor del documento: Agencia Digital de Andalucía

Tipo de documento: Guía

Fecha de elaboración: 19/02/2026

Fecha de última actualización: 10/03/2026

ÍNDICE DE CONTENIDOS

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

ÍNDICE DE TABLAS

1. Introducción

La incorporación progresiva de sistemas de Inteligencia Artificial (IA) en los servicios públicos de salud y en los servicios digitales de ciudades inteligentes (Smart Cities) ha transformado la naturaleza de los entornos operativos en los que se prestan servicios esenciales a la ciudadanía. Los sistemas de IA desplegados en estos ámbitos generan predicciones, clasificaciones, recomendaciones, priorizaciones y contenidos que influyen directamente en decisiones con impacto sobre la seguridad de las personas, la continuidad de los servicios públicos y el ejercicio de derechos fundamentales.

Desde la perspectiva de la ciberseguridad, estos sistemas presentan características que los diferencian de los sistemas de información convencionales. Un sistema de IA puede continuar funcionando desde un punto de vista técnico mientras la fiabilidad de los resultados que genera se ha visto comprometida. En estas circunstancias, el sistema puede seguir produciendo resultados aparentemente plausibles que influyen en procesos de decisión clínica, en la gestión de servicios urbanos o en otros procesos operativos, sin que el incidente sea inmediatamente detectable.

En esta guía se utiliza el término incidente de ciberseguridad con impacto algorítmico para referirse a aquellas situaciones en las que un sistema de IA continúa operativo mientras sus resultados han sido alterados o comprometidos. Este tipo de incidentes puede permanecer sin detectar durante períodos prolongados y afectar de forma acumulativa a decisiones y procesos que dependen del sistema.

Los sistemas de IA presentan además vectores de ataque específicos que no se corresponden plenamente con los incidentes clásicos de los sistemas de información convencionales, entre los que se incluyen el envenenamiento de datos, los ataques adversarios, la extracción o inversión de modelos, la manipulación de pipelines de aprendizaje automático o la inyección de instrucciones en sistemas generativos. El marco normativo vigente refuerza la necesidad de un enfoque especializado. El Reglamento (UE) 2024/1689, Reglamento de Inteligencia Artificial (RIA), la Directiva (UE) 2022/2555, el Reglamento General de Protección de Datos y el Esquema Nacional de Seguridad establecen, desde perspectivas complementarias, obligaciones de gestión de riesgos, notificación y respuesta que pueden converger sobre un mismo incidente. En particular, en el ámbito del RIA, la notificación de incidentes graves se regula en el artículo 73. La presente guía establece un marco operativo para la preparación, detección, análisis, contención, erradicación, recuperación y mejora continua ante incidentes de ciberseguridad que afecten a sistemas de IA en entornos de salud y smart cities.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

2. Objetivo y alcance

2.1. Objetivo

El objetivo de la presente guía es proporcionar un marco de referencia para la gestión de incidentes de ciberseguridad que afecten a sistemas de Inteligencia Artificial desplegados en entornos de salud y Smart Cities. La guía establece criterios y orientaciones para la preparación, detección, análisis, contención, erradicación, recuperación y mejora continua ante este tipo de incidentes, teniendo en cuenta las particularidades técnicas y operativas de los sistemas de IA.

Asimismo, la guía tiene por finalidad facilitar la identificación de incidentes que puedan comprometer la integridad o fiabilidad de los resultados generados por sistemas de IA, así como orientar la adopción de medidas que permitan reducir su impacto sobre los procesos y decisiones que dependen de dichos resultados.

El documento pretende servir como referencia para las organizaciones responsables de la operación, supervisión o uso de sistemas de IA en los ámbitos considerados, proporcionando un marco conceptual y operativo que complemente los procedimientos generales de gestión de incidentes de ciberseguridad.

2.2. Alcance

La presente guía se aplica a la gestión de incidentes de ciberseguridad que afecten a sistemas de Inteligencia Artificial utilizados en entornos de salud y en servicios digitales asociados a ciudades inteligentes.

A efectos de esta guía, se consideran dentro del alcance los incidentes que puedan afectar a los distintos componentes que integran un sistema de IA, incluyendo, entre otros, los modelos utilizados, los datos empleados para su entrenamiento o funcionamiento, los procesos de desarrollo y despliegue asociados al ciclo de vida del modelo, las interfaces de acceso al sistema y los resultados generados por el propio sistema.

La guía aborda incidentes que puedan comprometer la confidencialidad, integridad o disponibilidad de los sistemas de IA, así como aquellos que puedan afectar a la fiabilidad de los resultados producidos por dichos sistemas y, en consecuencia, a los procesos de decisión que dependen de ellos.

Quedan fuera del alcance de esta guía los incidentes que afecten exclusivamente a infraestructuras o sistemas de información que no incorporen componentes de Inteligencia Artificial, salvo en la medida en que dichos incidentes tengan un impacto directo sobre el funcionamiento o la seguridad de sistemas de IA incluidos en los ámbitos considerados.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

3. Referencias normativas

Las referencias incluidas en este apartado conforman el marco jurídico y técnico directamente aplicable a la gestión de incidentes de ciberseguridad en sistemas de Inteligencia Artificial desplegados en entornos de salud y Smart Cities. Se integran normas europeas, nacionales e internacionales sobre inteligencia artificial, protección de datos, ciberseguridad y seguridad de los sistemas de información del sector público.

- **Reglamento (UE) 2024/1689, Reglamento de Inteligencia Artificial (RIA):** Establece un marco jurídico armonizado para el desarrollo, la comercialización, la puesta en servicio y el uso de sistemas de IA en la Unión Europea. Introduce un enfoque basado en el riesgo y obligaciones específicas para proveedores y personas responsables del despliegue de sistemas de IA, incluyendo obligaciones de gestión de riesgos, documentación, trazabilidad, transparencia, supervisión humana y notificación de incidentes graves. En materia de incidentes graves, resulta especialmente relevante el artículo 73. Resulta especialmente aplicable a los sistemas de IA de alto riesgo desplegados en los ámbitos de salud y servicios públicos.

- **Reglamento (UE) 2016/679, Reglamento General de Protección de Datos (RGPD):** Garantiza la protección de las personas físicas respecto al tratamiento de datos personales. Establece las obligaciones de notificación de violaciones de seguridad de datos personales a la autoridad de control (art. 33) y a las personas afectadas (art. 34). Resulta plenamente aplicable al tratamiento de datos mediante sistemas de IA, con especial relevancia cuando el tratamiento involucra datos de categoría especial como datos de salud.

- **Ley Orgánica 3/2018, de Protección de Datos Personales y garantía de los derechos digitales (LOPDGDD):** Desarrolla y complementa el RGPD en el ordenamiento español, incorporando derechos digitales y mecanismos reforzados de transparencia, supervisión y rendición de cuentas.

- **Real Decreto 311/2022, por el que se regula el Esquema Nacional de Seguridad (ENS):** Define los principios y requisitos mínimos de seguridad aplicables a los sistemas de información del sector público español y de los proveedores tecnológicos que les prestan servicios. Regula la prevención, detección y respuesta a incidentes de seguridad, en particular en sus artículos 33 y 34, y resulta aplicable a los sistemas de IA operados en el ámbito del sector público.

- **Directiva (UE) 2022/2555, relativa a medidas para un elevado nivel común de ciberseguridad en la Unión (Directiva NIS2):** Refuerza, a nivel de la Unión, las obligaciones de gestión de riesgos, notificación de incidentes y supervisión para entidades esenciales e importantes, incluidas administraciones públicas y determinados proveedores tecnológicos. Establece plazos de notificación como la alerta temprana en 24 horas, la notificación en 72 horas y el informe final en un mes. Su aplicación en España debe entenderse junto con la normativa nacional de transposición y desarrollo que resulte vigente.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

- **Reglamento (UE) 2017/745, sobre los productos sanitarios (MDR) y Reglamento (UE) 2017/746, sobre los productos sanitarios para diagnóstico in vitro (IVDR):** Establecen los requisitos de seguridad y las obligaciones de vigilancia poscomercialización para productos sanitarios, incluyendo la notificación de incidentes graves. Resultan aplicables cuando un sistema de IA esté calificado como producto sanitario o producto sanitario para diagnóstico in vitro.

- **Reglamento (UE) 2025/327, sobre el Espacio Europeo de Datos Sanitarios (EHDS):** Establece un marco europeo para el acceso, uso e intercambio de datos electrónicos de salud. Resulta especialmente relevante cuando los sistemas de IA en el ámbito sanitario tratan datos de salud, se integran en infraestructuras de datos sanitarios o se despliegan en contextos sometidos a obligaciones específicas de interoperabilidad, acceso y gobernanza de datos sanitarios.

- **Reglamento (UE) 2024/2847, relativo a los requisitos horizontales de ciberseguridad para los productos con elementos digitales (Cyber Resilience Act, CRA):** Establece requisitos horizontales de ciberseguridad para productos con elementos digitales, incluyendo dispositivos, componentes y software que puedan incorporar funcionalidades de IA. Su relevancia práctica debe interpretarse conforme a su calendario de aplicación.

- **ISO/IEC 42001:2023, Information technology – Artificial intelligence – Management system**: Especifica los requisitos para implantar, mantener y mejorar un sistema de gestión de la IA, con el fin de garantizar un uso seguro, responsable y trazable de los sistemas de IA en las organizaciones.

- **ISO/IEC 23894:2023, Information technology – Artificial intelligence – Risk management:** Proporciona un marco estructurado para identificar, evaluar y tratar los riesgos asociados al desarrollo, despliegue y uso de la IA, incluyendo riesgos técnicos, organizativos, éticos y regulatorios.

- **ISO/IEC 27001:2022, Information security, cybersecurity and privacy protection – Information security management systems – Requirements:** Establece los requisitos para implantar y mantener un sistema de gestión de la seguridad de la información, aplicable a las organizaciones que operan sistemas y servicios de IA.

- **NIST AI Risk Management Framework 1.0 (2023**): Proporciona un marco de referencia para gestionar los riesgos de la IA en todo su ciclo de vida. Se considera una referencia técnica complementaria cuando su utilización resulte compatible con el marco jurídico europeo.

- **Guías CCN-STIC del Centro Criptológico Nacional:** En particular, la CCN-STIC 817 sobre gestión de ciberincidentes, que establece los procedimientos de notificación y gestión de incidentes en el ámbito del ENS, y otras guías CCN-STIC que resulten aplicables a la seguridad de los sistemas de información que soportan los sistemas de IA.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

4. Definiciones

Este capítulo proporciona una tabla clara y precisa de acrónimos y términos relevantes utilizados a lo largo del documento.

Acrónimos/Términos	Definición
AEMPS	Agencia Española de Medicamentos y Productos Sanitarios. Autoridad competente en materia de vigilancia de productos sanitarios en España.
AEPD	Agencia Española de Protección de Datos. Autoridad de control independiente encargada de velar por el cumplimiento de la normativa de protección de datos personales en España.
AESIA	Agencia Española de Supervisión de la Inteligencia Artificial. Autoridad nacional de supervisión en materia de inteligencia artificial.
Ajuste fino (Fine-tuning)	Proceso mediante el cual se adapta un modelo previamente entrenado utilizando datos más específicos, con el objetivo de mejorar su funcionamiento en un contexto o caso de uso concreto.
API (Interfaz de Programación de Aplicaciones)	Mecanismo que permite que distintos sistemas o aplicaciones se comuniquen entre sí de forma controlada, intercambiando datos o funcionalidades.
Ataque adversario	Perturbación calculada de las entradas de un modelo de IA, diseñada para inducir un error específico en la salida del modelo sin que la alteración sea perceptible o significativa para un observador humano.
Baseline del modelo	Conjunto de métricas de rendimiento del modelo documentadas y verificadas en el momento de su validación y despliegue, que sirven como referencia para la detección de degradaciones o alteraciones en su comportamiento durante la operación.
Cadena de custodia	Conjunto de procedimientos que garantizan la preservación, integridad y trazabilidad de las evidencias desde el momento en que se recogen hasta su uso final.
Cadena de suministro de IA	Conjunto de componentes, librerías, frameworks, modelos preentrenados, datos y servicios de terceros que se integran en el desarrollo, entrenamiento, validación o despliegue de un sistema de IA.
CCN-CERT	Centro Criptológico Nacional - Computer Emergency Response Team. Equipo de respuesta ante incidentes de seguridad de la información del Centro Criptológico Nacional, adscrito al Centro Nacional de Inteligencia. Actúa como CSIRT de referencia para el sector público español.
Concept drift	Cambio en la relación estadística entre las variables de entrada y las salidas esperadas del modelo, producido por la evolución del contexto operativo a lo largo del tiempo.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

Acrónimos/Términos	Definición
CSIRT (Computer Security Incident Response Team)	Equipo de respuesta ante incidentes de seguridad informática. Organización o equipo responsable de recibir, analizar y responder a incidentes de ciberseguridad.
Data drift	Cambio en la distribución estadística de los datos de entrada que recibe el modelo en producción respecto a la distribución de los datos con los que fue entrenado.
Denegación de servicio	Ataque o incidente de seguridad cuyo objetivo es impedir o degradar el acceso normal a un sistema, servicio o aplicación, provocando que las personas usuarias legítimas no puedan utilizarlo, normalmente mediante la saturación de recursos o el envío masivo de peticiones.
ENISA (Agencia de la Unión Europea para la Ciberseguridad)	Es el organismo europeo encargado de apoyar y reforzar la ciberseguridad en los Estados miembros, mediante la elaboración de guías, recomendaciones, marcos de referencia y buenas prácticas en materia de seguridad de la información y tecnologías digitales.
Envenenamiento de datos	Ataque consistente en la inserción, modificación o eliminación deliberada de datos en los conjuntos utilizados para entrenar o ajustar un modelo de IA, con el objetivo de alterar su comportamiento.
Extracción de modelo	Ataque mediante el cual un adversario reconstruye o aproxima un modelo de IA a partir de consultas sistemáticas a su interfaz de inferencia, obteniendo una réplica funcional del modelo.
Guardrail	Mecanismo de control implementado en un sistema de IA generativa para restringir su comportamiento dentro de los límites definidos por la organización, incluyendo filtros de contenido, restricciones temáticas y mecanismos de detección de inyecciones.
Hash	Valor generado a partir de unos datos mediante una función matemática que los transforma en una cadena de tamaño fijo. Se utiliza para verificar la integridad de la información, ya que cualquier cambio en los datos originales produce un hash distinto.
Hiperparámetro	Parámetro de configuración que se establece antes de entrenar un modelo y que controla cómo se realiza el aprendizaje. No se aprende automáticamente a partir de los datos, sino que se define previamente para influir en el comportamiento y rendimiento del modelo.
Impacto algorítmico	Consecuencia de un incidente de ciberseguridad que afecta a la fiabilidad de los resultados generados por un sistema de IA, pudiendo influir en decisiones clínicas, de gestión urbana o de prestación de servicios sin que se produzca una interrupción técnica del sistema.
Inferencia	Proceso mediante el cual un modelo de IA genera un resultado (predicción, clasificación, recomendación, contenido) a partir de unos datos de entrada.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

Acrónimos/Términos	Definición
Inversión de modelo	Ataque mediante el cual se infiere información sobre los datos utilizados para entrenar un modelo a partir del análisis de sus salidas.
Inyección de instrucciones (prompt injection)	Ataque contra sistemas de IA generativa consistente en introducir instrucciones maliciosas en las entradas del sistema con el objetivo de alterar su comportamiento, evadir sus controles de seguridad o acceder a información restringida.
LLM (Large Language Model)	Modelo de lenguaje de gran tamaño. Tipo de modelo fundacional entrenado sobre grandes volúmenes de texto, capaz de generar, comprender y procesar lenguaje natural.
Log	Registro automático de eventos que recoge información sobre las acciones, operaciones o incidencias que se producen en un sistema, aplicación o servicio.
MLOps	Conjunto de prácticas, procesos y herramientas para el desarrollo, entrenamiento, validación, despliegue, monitorización y mantenimiento de modelos de aprendizaje automático en entornos de producción.
Modelo	En el contexto de esta guía, conjunto de parámetros, arquitectura y configuración que definen el comportamiento de un sistema de IA entrenado para realizar una tarea específica.
Modelo fundacional	Modelo de IA de gran escala, generalmente entrenado sobre grandes volúmenes de datos, que sirve como base para su adaptación a tareas específicas. Los modelos de lenguaje de gran tamaño (LLM) son un ejemplo de modelo fundacional.
Modelo propietario	Modelo desarrollado y controlado por una organización concreta, cuyo código, funcionamiento interno y datos no son públicos. Su uso está sujeto a licencias, condiciones comerciales y restricciones de acceso definidas por el propietario.
Modo supervisado	Modalidad de operación en la que el sistema funciona de manera normal, pero todos sus resultados son revisados y validados por una persona antes de incorporarse al proceso de decisión.
Pipeline	Secuencia automatizada de procesos que abarca las etapas de preparación de datos, entrenamiento, validación, despliegue y actualización de un modelo de IA.
Prediction drift	Cambio estadísticamente significativo en la distribución de las salidas generadas por un modelo de IA en producción respecto a la distribución esperada de dichas salidas.
Procedimiento alternativo	Proceso documentado y operativo que permite mantener la continuidad de un proceso o servicio que depende de un sistema de IA cuando dicho sistema no está disponible o sus resultados no son fiables.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

Acrónimos/Términos	Definición
Prompt del sistema	Instrucción o conjunto de instrucciones predefinidas que configuran el comportamiento de un sistema de IA generativa, estableciendo su rol, sus restricciones y sus pautas de respuesta.
Puerta trasera (backdoor)	Comportamiento oculto introducido deliberadamente en un modelo de IA, que se activa únicamente cuando el modelo recibe una entrada que contiene un patrón específico (trigger), produciendo un resultado predeterminado por el atacante.
RIA	La referencia abreviada al Reglamento (UE) 2024/1689 de inteligencia artificial.
Servidor de Comando y Control (C&C)	Infraestructura utilizada por los atacantes para controlar dispositivos infectados y gestionar redes de malware, como botnets.
Shadow AI	Uso de herramientas o sistemas de IA dentro de una organización sin la autorización, el conocimiento o la supervisión de las personas responsables de seguridad y gobernanza de la organización.
SIEM	Es una solución de seguridad que centraliza y analiza la información y los eventos generados por los sistemas (logs), con el objetivo de detectar incidentes de seguridad, comportamientos anómalos y facilitar la supervisión y el cumplimiento normativo.
SOC (Security Operations Center)	Centro de operaciones de seguridad. Unidad organizativa responsable de la monitorización continua, la detección y la respuesta inicial ante eventos e incidentes de ciberseguridad.
Técnicas de persistencia	Conjunto de métodos utilizados por un atacante para mantener el acceso a un sistema de forma prolongada, incluso después de reinicios, cierres de sesión o intentos de limpieza.
Token	En el contexto de autenticación y control de acceso, credencial o artefacto digital que acredita la identidad, los permisos o la sesión de una persona usuaria o sistema para acceder a un recurso.
Trigger	Patrón específico de entrada diseñado para activar una puerta trasera introducida en un modelo de IA, provocando un comportamiento predeterminado por el atacante.
Vector de ataque	Método o vía utilizada por un atacante para acceder a un sistema, comprometer un componente o explotar una vulnerabilidad.

TABLA 1: DEFINICIONES EMPLEADAS.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

5. Particularidades de los incidentes de ciberseguridad en sistemas de IA

Los incidentes de ciberseguridad que afectan a sistemas de Inteligencia Artificial presentan características que los diferencian de los incidentes que afectan a sistemas de información convencionales. Estas diferencias influyen tanto en la forma en que los incidentes se manifiestan como en los componentes que pueden verse comprometidos. También afectan a las fases del proceso de respuesta y al tipo de impacto que pueden generar en los entornos de salud y Smart Cities.

5.1. Diferencias frente a incidentes de ciberseguridad convencionales

Los incidentes de ciberseguridad en sistemas de Inteligencia Artificial presentan diferencias estructurales respecto de los que afectan a sistemas de información convencionales, tanto en la forma en que se manifiestan como en la forma en que se detectan, se analizan y se gestionan.

La tabla siguiente presenta una síntesis de las principales diferencias operativas entre los incidentes que afectan a sistemas de información convencionales y aquellos que afectan a sistemas de Inteligencia Artificial:

Características	Incidente en sistema de información convencional	Incidente en sistema de IA
Manifestación	Visible: caída del servicio, errores, alertas de seguridad	Puede no ser visible: el sistema continúa funcionando mientras produce resultados alterados
Detección	Monitorización de infraestructura y eventos de seguridad	Puede requerir análisis del comportamiento del modelo y de las salidas generadas
Componentes afectados	Infraestructura, datos almacenados, configuraciones, aplicaciones	Además de la infraestructura: modelo, datos de entrenamiento, pipeline de desarrollo y despliegue, lógica inferencial
Naturaleza del impacto	Impacto técnico directo sobre sistemas o datos	Puede producir impacto algorítmico: decisiones incorrectas basadas en resultados alterados
Persistencia	Resoluble mediante restauración de sistemas o datos	Un modelo comprometido puede seguir produciendo resultados incorrectos tras la eliminación del vector de ataque
Alcance temporal	Acotable al período de compromiso	Las decisiones adoptadas durante el período afectado pueden tener efectos posteriores

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

Erradicación	Eliminación de código malicioso, cierre de vulnerabilidades, restauración desde copias de seguridad	Puede requerir reentrenamiento del modelo, revisión de datos y reconstrucción de pipelines

TABLA 2: DIFERENCIAS ENTRE INCIDENTES CONVENCIONALES E INCIDENTES EN SISTEMAS DE IA

5.2. Componentes vulnerables de un sistema de IA

Un sistema de IA en producción comprende múltiples componentes, cada uno de los cuales puede ser objeto de un incidente de ciberseguridad con consecuencias diferenciadas. La identificación de estos componentes es necesaria para determinar el alcance de un incidente y orientar las acciones de respuesta.

Componente	Descripción	Ejemplos de incidentes
Datos de entrenamiento	Datos utilizados para entrenar, validar o ajustar el modelo	Envenenamiento de datos, manipulación, exfiltración
Modelo	Parámetros, arquitectura y configuración del modelo entrenado	Manipulación del modelo, sustitución, extracción o inversión
Pipeline de desarrollo y despliegue	Procesos de entrenamiento, validación, despliegue y actualización del modelo	Inyección de código, despliegue de modelos no autorizados, sabotaje del pipeline
Interfaz de inferencia	Interfaz mediante la cual el sistema recibe consultas y genera resultados	Explotación de la API, denegación de servicio, consultas masivas para extracción
Instrucciones y plantillas	Configuración e instrucciones utilizadas por sistemas generativos	Inyección de instrucciones, manipulación de configuraciones
Datos de inferencia	Datos introducidos en el sistema durante su operación	Ataques adversarios, manipulación de entradas
Resultados	Salidas generadas por el sistema e integradas en procesos de decisión	Manipulación en postprocesamiento, alteración de la presentación de resultado

TABLA 3: COMPONENTES VULNERABLES DE UN SISTEMA DE IA

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

5.3. Implicaciones para las fases de gestión de incidentes

Las diferencias y los componentes descritos en los apartados anteriores tienen implicaciones concretas sobre cada una de las fases del proceso de respuesta a incidentes.

I. Detección:

La identificación de un incidente en un sistema de IA no puede basarse exclusivamente en indicadores convencionales como la interrupción del servicio, la alteración de registros o los accesos no autorizados. Un incidente con impacto algorítmico puede no generar alertas en los sistemas de monitorización de infraestructura. La detección requiere incorporar indicadores específicos relativos al comportamiento del modelo, tales como variaciones en las distribuciones de las salidas, cambios en las métricas de rendimiento, divergencias entre los resultados del sistema y los resultados esperados según datos de referencia, o alteraciones en las distribuciones de los datos de entrada.

II. Análisis:

Determinar el origen y el alcance de un incidente en un sistema de IA presenta una complejidad específica. La relación entre los datos de entrada, los parámetros del modelo y los resultados generados no es directamente trazable en la mayoría de los modelos. Un resultado comprometido puede tener su origen en la manipulación de los datos de entrenamiento, en la alteración del modelo desplegado, en la modificación de los datos de entrada durante la inferencia o en una combinación de estos factores. El análisis del incidente requiere capacidades técnicas específicas para examinar los componentes del sistema de IA de forma diferenciada.

III. Contención:

Las medidas de contención en sistemas de IA deben tener en cuenta que la retirada inmediata del sistema puede interrumpir procesos asistenciales o servicios urbanos que dependen de sus resultados. La contención puede requerir la activación de modelos de respaldo, la reversión a versiones anteriores del modelo o la incorporación de mecanismos de supervisión humana sobre las salidas del sistema, además de las medidas de contención convencionales sobre la infraestructura.

IV. Erradicación:

La eliminación de la causa raíz en un sistema de IA puede implicar el reentrenamiento del modelo con datos verificados, la sustitución de componentes del pipeline de desarrollo, la revisión de los conjuntos de datos utilizados o la eliminación de dependencias comprometidas en la cadena de suministro. Estas acciones requieren plazos de ejecución más amplios que las medidas de erradicación en sistemas convencionales y la participación de perfiles técnicos especializados en aprendizaje automático.

V. Evaluación del impacto:

El impacto de un incidente en un sistema de IA no se limita a las dimensiones de confidencialidad, integridad y disponibilidad de los activos tecnológicos. Debe evaluarse también el impacto sobre la fiabilidad de las decisiones adoptadas a partir de los resultados del

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

sistema durante el período en que el incidente estuvo activo. En entornos de salud, esto puede implicar la revisión de decisiones clínicas. En entornos de Smart Cities, puede implicar la revisión de actuaciones sobre servicios públicos.

5.4. Impacto potencial en entornos de salud y Smart Cities

Las particularidades descritas en los apartados anteriores adquieren una dimensión específica en los entornos de salud y Smart Cities debido a la criticidad de los procesos en los que intervienen los sistemas de IA.

En entornos de salud, un incidente de ciberseguridad en un sistema de IA puede dar lugar a decisiones clínicas fundamentadas en resultados comprometidos, retrasos en la atención derivados de la indisponibilidad de sistemas integrados en flujos asistenciales, exposición de datos de salud, introducción o amplificación de sesgos con impacto sobre la equidad asistencial y pérdida de confianza del personal sanitario y de las personas pacientes en los sistemas clínicos.

En entornos de Smart Cities, un incidente puede dar lugar a decisiones erróneas en la gestión urbana, asignación incorrecta de recursos, información inadecuada dirigida a la ciudadanía, degradación de la calidad del servicio público, afectación de la seguridad pública cuando los resultados del sistema de IA se integran en procesos de gestión de emergencias o movilidad, y discriminación algorítmica en la prestación de servicios municipales.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

6. Tipología de incidentes de ciberseguridad en sistemas de IA

Los incidentes de ciberseguridad que afectan a sistemas de Inteligencia Artificial pueden manifestarse de formas diversas en función del componente del sistema que resulte afectado. A partir de los componentes vulnerables identificados en el apartado 5.2, la presente sección establece una tipología de incidentes organizada en cuatro categorías que permite orientar la clasificación inicial del incidente durante las primeras fases del proceso de respuesta.

La siguiente tabla resume las categorías de incidentes tratadas en este capítulo:

Categoría de incidente	Componente principal afectado	Ejemplos representativos
Incidentes relacionados con los datos	Datos de entrenamiento o inferencia	Exfiltración de datos, envenenamiento, manipulación de entradas
Incidentes relacionados con los modelos	Modelo desplegado o artefacto del modelo	Manipulación del modelo, sustitución, extracción, inversión
Incidentes relacionados con el ciclo de vida del modelo	Entrenamiento, validación, despliegue y actualización	Alteración del pipeline, validación comprometida, cadena de suministro
Incidentes relacionados con el uso del sistema	Operación del sistema e interacción con personas usuarias o sistemas externos	Ataques adversarios, inyección de instrucciones, Shadow AI

TABLA 4: TIPOLOGÍA GENERAL DE INCIDENTES DE CIBERSEGURIDAD EN SISTEMAS IA

6.1. Incidentes relacionados con los datos

Los datos constituyen un elemento esencial en el funcionamiento de los sistemas de IA. Los incidentes que afectan a los datos pueden alterar el comportamiento del modelo, comprometer la confidencialidad de la información utilizada o degradar la fiabilidad de los resultados generados durante la operación del sistema. En función del momento del ciclo de vida en que se produzcan, estos incidentes pueden afectar de forma persistente al comportamiento del modelo o provocar errores asociados a casos concretos durante su utilización.

Una forma particularmente relevante de incidente en esta categoría es el envenenamiento de datos, mediante el cual se introducen datos manipulados en los conjuntos utilizados para entrenar o ajustar el modelo con el objetivo de alterar su comportamiento. También debe considerarse la manipulación de datos de inferencia, que puede dar lugar a resultados incorrectos aun cuando el modelo desplegado no haya sido modificado.

Tipo de incidente	Descripción	Impacto	Ejemplo

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

Exfiltración de datos de entrenamiento	Acceso no autorizado y extracción de los conjuntos de datos utilizados para entrenar el modelo	Violación de seguridad de datos personales, exposición de información sensible y facilitación de ataques posteriores	Exfiltración de un dataset clínico utilizado para entrenar un modelo de diagnóstico
Envenenamiento de datos de entrenamiento	Inserción, modificación o eliminación deliberada de datos en los conjuntos de entrenamiento, incluyendo la manipulación gradual de datos en ciclos de reentrenamiento continuo	Alteración del comportamiento del modelo, introducción de sesgos o puertas traseras, impacto persistente	Inserción de registros manipulados en un dataset de urgencias para sesgar la priorización
Corrupción de datos de inferencia	Alteración de los datos de entrada que el sistema recibe en producción	Resultados incorrectos en las inferencias afectadas y decisiones erróneas asociadas	Manipulación de datos que alimentan un modelo de predicción ambiental
Fuga de datos a través de resultados del modelo	Los resultados revelan información sobre datos de entrenamiento o de entrada de otras personas usuarias	Violación de la privacidad y riesgo de reidentificación	Un modelo de lenguaje sanitario que reproduce fragmentos de historiales clínicos

TABLA 5: INCIDENTES RELACIONADOS CON LOS DATOS

6.2. Incidentes relacionados con los modelos

Los modelos de Inteligencia Artificial constituyen el núcleo funcional del sistema. Los incidentes que afectan directamente a los modelos pueden alterar su funcionamiento, comprometer su confidencialidad o reducir la fiabilidad de los resultados que generan. El compromiso de un modelo tiene implicaciones particularmente graves debido a la persistencia potencial de sus efectos y a la dificultad de detección.

Tipo de incidente	Descripción	Impacto	Ejemplo
Manipulación de parámetros del modelo	Modificación no autorizada de los pesos, sesgos u otros parámetros del modelo desplegado	Alteración controlada del comportamiento del modelo	Modificación de un modelo de priorización para infravalorar determinados incidentes

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

Sustitución del modelo	Reemplazo del modelo legítimo por una versión manipulada o distinta	Control del comportamiento del sistema y pérdida de confianza en sus resultados	Sustitución de un modelo clínico durante una actualización
Extracción del modelo	Reconstrucción del modelo interno mediante consultas sistemáticas	Pérdida de propiedad intelectual y facilitación de ataques posteriores	Consultas masivas a una interfaz de inferencia para replicar el comportamiento del modelo
Inversión del modelo	Inferencia de información sobre los datos de entrenamiento a partir de las salidas	Exposición de datos personales utilizados en el entrenamiento	Reconstrucción parcial de características de personas pacientes a partir de respuestas del modelo

TABLA 6: INCIDENTES RELACIONADOS CON LOS MODELOS

6.3. Incidentes relacionados con el ciclo de vida del modelo

Los sistemas de IA se desarrollan y mantienen mediante procesos que incluyen el entrenamiento, la validación, el despliegue y la actualización periódica de los modelos. Estos procesos forman parte del ciclo de vida del modelo y constituyen un ámbito específico de exposición a incidentes de ciberseguridad.

Dentro de esta categoría se incluyen también los incidentes originados en la cadena de suministro de componentes de IA: librerías, frameworks, modelos preentrenados descargados de repositorios públicos, contenedores e imágenes de despliegue. El compromiso de cualquiera de estos elementos puede introducir vulnerabilidades o puertas traseras en el sistema sin que la organización lo perciba, dado que el componente comprometido se integra en el proceso legítimo de construcción o actualización del sistema.

Tipo de incidente	Descripción	Impacto	Ejemplo
Manipulación del proceso de entrenamiento	Alteración del código, configuraciones o hiperparámetros del entrenamiento	Modelo resultante con comportamiento alterado	Modificación de scripts de entrenamiento para introducir un sesgo sistemático
Modificación de datos en el pipeline	Alteración de los datos a lo largo del procesamiento	Envenenamiento efectivo sin alterar el dataset original	Inyección de datos manipulados en una fase de preprocesamiento

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

Alteración de procesos de validación	Manipulación de métricas, conjuntos de test o umbrales	Un modelo comprometido supera las validaciones y se despliega	Modificación del conjunto de test para ocultar el sesgo introducido
Despliegue de modelos no autorizados	Inserción de un modelo no aprobado en el pipeline de despliegue	Modelo no validado en producción, con comportamiento desconocido o malicioso	Sustitución del artefacto del modelo durante el despliegue automatizado
Compromiso de la cadena de suministro de IA	Incorporación de librerías, herramientas, contenedores o modelos preentrenados comprometidos en el pipeline de desarrollo o despliegue	Introducción de vulnerabilidades o puertas traseras difíciles de detectar	Uso de un modelo preentrenado descargado de un repositorio externo que incorpora una puerta trasera

TABLA 7: INCIDENTES RELACIONADOS CON EL CICLO DE VIDA DEL MODELO

6.4. Incidentes relacionados con el uso del sistema

Los incidentes relacionados con el uso del sistema pueden producirse durante la operación del sistema como consecuencia de interacciones maliciosas o manipuladas con sus entradas, sus salidas o los procesos que dependen de ellas. En esta categoría se incluyen tanto incidentes asociados a sistemas predictivos o clasificatorios como incidentes específicos de sistemas de IA generativa.

Tipo de incidente	Descripción	Impacto	Ejemplo
Ataque adversario sobre entradas	Perturbación calculada de las entradas para inducir un error específico en la salida	Resultados incorrectos para entradas manipuladas	Modificación imperceptible de una imagen médica que provoca una clasificación errónea
Manipulación de salidas	Alteración de los resultados después de que el modelo los genera	Los resultados presentados no se corresponden con los generados por el modelo	Alteración del módulo que traduce puntuaciones en categorías de prioridad clínica
Interferencia con decisiones automatizadas	Manipulación de los mecanismos que aplican automáticamente los resultados	Acciones automatizadas incorrectas	Alteración de umbrales que activan alertas en un sistema de gestión de emergencias

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

Utilización de IA no autorizada (Shadow AI)	Uso dentro de la organización de herramientas de IA no autorizadas ni supervisadas	Información tratada fuera de los mecanismos de gobernanza, pérdida de trazabilidad, resultados no controlados	Personal sanitario que utiliza un modelo de lenguaje externo para procesar datos clínicos
Inyección de instrucciones maliciosas	Introducción de instrucciones destinadas a alterar el comportamiento de un sistema de IA generativa	Generación de respuestas no previstas, exposición de información o elusión de restricciones	Inserción de instrucciones ocultas en una consulta para alterar el comportamiento de un asistente generativo

TABLA 8: INCIDENTES RELACIONADOS CON EL USO DEL SISTEMA

7. Principios de respuesta a incidentes en sistemas IA

La respuesta a incidentes de ciberseguridad en sistemas de Inteligencia Artificial debe regirse por un conjunto de principios operativos que orienten la toma de decisiones durante todas las fases del proceso de gestión del incidente. Estos principios complementan los principios generales de respuesta a incidentes de ciberseguridad con consideraciones específicas derivadas de la naturaleza de los sistemas de IA y del impacto que sus resultados pueden tener sobre las personas y los servicios.

I. Seguridad de las personas como prioridad

Toda decisión adoptada durante la respuesta a un incidente en un sistema de IA debe evaluarse, en primer lugar, desde la perspectiva de su impacto sobre la seguridad de las personas. En entornos de salud, la seguridad de la persona paciente debe prevalecer sobre la continuidad del sistema. En entornos de Smart Cities, la seguridad pública debe prevalecer sobre la operatividad del servicio cuando los resultados generados por el sistema de IA puedan provocar un perjuicio relevante.

II. Supervisión humana efectiva

Durante la gestión de un incidente, los mecanismos de supervisión humana sobre los resultados generados por el sistema de IA deben reforzarse. Cuando exista incertidumbre sobre la fiabilidad del sistema o sobre la integridad de los resultados generados, las decisiones que dependan de dichos resultados deben ser objeto de revisión humana o de validación adicional.

III. Contención temprana

La contención debe constituir una prioridad una vez confirmado un incidente que afecte a un sistema de IA. Dado que los incidentes con impacto algorítmico pueden producir daños acumulativos mientras el sistema continúa generando resultados, cualquier demora injustificada en la adopción de medidas de contención puede amplificar las consecuencias del incidente.

IV. Evaluación del impacto algorítmico

El impacto de un incidente en un sistema de IA no se limita necesariamente a la indisponibilidad del sistema o al acceso no autorizado a datos. Debe evaluarse si el comportamiento del modelo se ha visto alterado y en qué medida los resultados generados durante el período afectado pueden haber influido en decisiones operativas, clínicas o de gestión urbana.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

V. Preservación de evidencias

Las acciones adoptadas durante la respuesta al incidente deben ejecutarse de forma que se preserve la integridad de las evidencias necesarias para el análisis posterior. En sistemas de IA, estas evidencias pueden incluir versiones del modelo, datos de entrenamiento, registros de inferencia, métricas de rendimiento, configuraciones del pipeline de desarrollo y registros de actividad del sistema.

VI. Trazabilidad completa

Todas las acciones adoptadas durante la gestión del incidente, así como las decisiones tomadas y los hallazgos obtenidos durante el análisis, deben documentarse de forma sistemática. La trazabilidad de las actuaciones realizadas facilita la comprensión del incidente, permite reconstruir su evolución y contribuye a la mejora de los mecanismos de respuesta.

VII. Comunicación coordinada

La gestión de incidentes que afectan a sistemas de IA puede requerir la intervención de múltiples perfiles profesionales y puede implicar obligaciones de notificación ante diferentes autoridades. La comunicación durante el incidente debe ser coordinada, coherente y oportuna, tanto a nivel interno entre los equipos implicados como en las comunicaciones externas que resulten necesarias.

VIII. Proporcionalidad de la respuesta

Las medidas adoptadas durante la respuesta deben ser proporcionales a la severidad del incidente, al impacto real o potencial sobre las personas y los servicios, y a las capacidades de la organización. Este principio implica adaptar la intensidad de las medidas de respuesta al nivel de riesgo identificado.

IX. Aprendizaje organizativo

Todo incidente constituye una fuente de información para la mejora de la seguridad de los sistemas de IA y de la capacidad de respuesta de la organización. Las lecciones aprendidas a partir del análisis del incidente deben incorporarse a los procesos de gestión de riesgos, a los mecanismos de detección y a los procedimientos de respuesta, con el fin de reducir la probabilidad de incidentes similares en el futuro.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

8. Modelo organizativo de respuesta a incidentes en sistemas de IA

La respuesta a incidentes de ciberseguridad en sistemas de IA requiere la participación coordinada de perfiles que, en la gestión de incidentes convencionales, no siempre intervienen de forma estructurada. El análisis de un modelo comprometido no puede realizarse únicamente desde la perspectiva de infraestructura, la evaluación del impacto clínico o sobre un servicio público no puede realizarse sin las personas responsables funcionales del proceso afectado, y la determinación de las obligaciones de notificación puede requerir la intervención simultánea del Delegado de Protección de Datos, de la persona responsable de la gobernanza de IA y de la asesoría jurídica.

El modelo organizativo de respuesta debe estar definido antes de que se produzca el incidente.

8.1. Roles y responsabilidades

Rol	Responsabilidades en la respuesta
Persona coordinadora del incidente	Dirección operativa de la respuesta, movilización de recursos, convocatorias de situación, supervisión de plazos, autorización de acciones críticas
Equipo de ciberseguridad (CSIRT/SOC)	Detección y triaje inicial, análisis forense, contención y erradicación sobre infraestructura, cadena de custodia de evidencias
Equipo de IA y ciencia de datos	Análisis del modelo, datos y pipelines; evaluación de la fiabilidad de los resultados; reentrenamiento, restauración y validación del modelo
Persona responsable de procesos asistenciales	Evaluación del impacto clínico, activación de procedimientos alternativos, revisión de decisiones afectadas
Persona responsable de servicios urbanos	Evaluación del impacto sobre servicios públicos, activación de procedimientos alternativos, coordinación operativa
Delegado/a de Protección de Datos	Evaluación de brecha de datos personales, determinación de notificaciones RGPD, asesoramiento sobre tratamiento de datos
Persona responsable de gobernanza de IA	Evaluación de implicaciones regulatorias del RIA, coordinación con proveedor, integración con gestión de riesgos del sistema
Dirección	Decisiones estratégicas, autorización de comunicaciones públicas, asignación de recursos extraordinarios
Proveedor del sistema de IA	Colaboración técnica, análisis de componentes bajo su responsabilidad, desarrollo de correcciones conforme a contrato

TABLA 9: ROLES Y RESPONSABILIDADES EN LA RESPUESTA A INCIDENTES EN SISTEMAS IA

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

La composición concreta del equipo de respuesta dependerá de la estructura y capacidades de cada organización. No todas las organizaciones dispondrán de todos los perfiles indicados de forma interna. En estos casos, los acuerdos contractuales con proveedores y las capacidades de los equipos de respuesta nacionales de referencia cobran especial relevancia.

8.2. Mecanismos de coordinación

La respuesta debe contar con los siguientes mecanismos, definidos y operativos con carácter previo al incidente:

- Reuniones de situación con periodicidad adaptada a la severidad del incidente.
- Canal de comunicación seguro compartido por los miembros del equipo de respuesta, preferiblemente independiente de la infraestructura potencialmente comprometida.
- Documentación centralizada de acciones, decisiones, hallazgos y comunicaciones.
- Punto de coordinación de notificaciones que centralice la gestión de las comunicaciones a autoridades externas.
- Criterios de escalado definidos que permitan elevar la gestión del incidente cuando concurran circunstancias como la afectación a la seguridad de las personas, el compromiso de datos de categoría especial, la sospecha de alteración de la fiabilidad del modelo o la necesidad de activar notificaciones regulatorias.

8.3. Autoridad para la suspensión de sistemas de IA

El plan de respuesta debe establecer con claridad quién tiene la autoridad para decidir la suspensión de un sistema de IA en producción. Se recomienda que:

- La persona coordinadora del incidente tenga autoridad para la suspensión inmediata cuando exista riesgo para la seguridad de las personas.
- En los demás casos, la suspensión se decida conjuntamente por la persona coordinadora del incidente, la persona responsable del proceso afectado y la persona responsable de gobernanza de IA.
- La decisión y su justificación queden documentadas de forma inmediata.

8.4. Integración con la gobernanza del sistema de IA

El modelo organizativo de respuesta no debe operar de forma aislada, sino integrarse con el marco de gobernanza del sistema de IA establecido por la organización y como se define en la *Guía de modelo de gobierno de IA.* La gestión del incidente debe poder apoyarse en la documentación del sistema, la evaluación de riesgos vigente, los registros del ciclo de vida del modelo y los mecanismos de supervisión ya definidos. Esta integración es especialmente relevante en sistemas cuya operación dependa de proveedores externos, de procesos de reentrenamiento periódico o de componentes reutilizados, donde la claridad previa sobre responsabilidades y mecanismos de control condiciona directamente la eficacia de la respuesta.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

9. Proceso de respuesta a incidentes en sistemas de IA

El proceso de respuesta se estructura en las fases que se desarrollan a continuación. Estas fases no sustituyen los procedimientos generales de gestión de incidentes de ciberseguridad de la organización, sino que los complementan con las especificidades derivadas de la naturaleza, los componentes y los riesgos propios de los sistemas de IA descritos en las secciones 5 y 6 de esta guía. Los roles y mecanismos de coordinación referidos a lo largo de esta sección son los definidos en la sección 8.

9.1. Preparación

La preparación constituye el fundamento sobre el que se sustenta la capacidad real de una organización para responder de forma eficaz a un incidente de ciberseguridad en un sistema de IA. A diferencia de las fases posteriores, que se activan cuando el incidente se ha producido o se sospecha, la preparación es una actividad continua que debe realizarse con carácter previo y mantenerse actualizada a lo largo del tiempo.

Las actividades de preparación descritas en esta sección complementan las capacidades generales de respuesta a incidentes de la organización con las especificidades propias de los sistemas de IA.

9.1.1. Inventario y clasificación de sistemas de IA

La organización debe mantener un inventario actualizado y accesible de todos los sistemas de IA que se encuentren en fase de operación. Este inventario sirve como recurso operativo que cumple una función directa durante la respuesta a incidentes por lo que se permite identificar de forma inmediata qué sistema está afectado, qué procesos de decisión dependen de sus resultados, qué datos trata y de qué categoría son, quién es la persona responsable de su supervisión, qué procedimientos alternativos existen y cuál es el marco regulatorio aplicable.

El inventario debe incluir, para cada sistema de IA en operación, como mínimo, los elementos que se recogen en la siguiente tabla:

Categoría	Elementos
Identificación	Nombre del sistema, versión en producción, proveedor o desarrollador, fecha de despliegue inicial, fecha de última actualización o modificación relevante.
Clasificación regulatoria	Nivel de riesgo conforme al RIA (alto riesgo, riesgo limitado, riesgo mínimo), calificación como producto sanitario o software sanitario conforme al MDR o IVDR en su caso, sujeción a requisitos específicos del EHDS cuando proceda.

Criticidad operativa	Procesos que dependen de los resultados del sistema, nivel de criticidad de dichos procesos para la organización y para las personas, existencia de procedimientos alternativos documentados y operativos que permitan mantener el proceso sin el sistema de IA, tiempo estimado necesario para la activación del procedimiento alternativo y limitaciones conocidas del mismo.
Componentes técnicos	Tipo de modelo (clasificación, regresión, generativo, etc.), ubicación de los datos de entrenamiento, validación e inferencia, descripción del pipeline de MLOps, APIs de inferencia, componentes de soporte necesarios para la operación del sistema.
Datos tratados	Tipología de los datos utilizados por el sistema, existencia de datos de categoría especial conforme al RGPD (datos de salud, datos biométricos, etc.), bases jurídicas del tratamiento, volumen aproximado de datos procesados.
Responsabilidades	Responsable de la supervisión operativa del sistema, equipo de IA o ciencia de datos asignado, Delegado de Protección de Datos, responsable funcional del proceso en el que se integra el sistema, datos de contacto del proveedor para incidentes.
Acuerdos contractuales	Cláusulas específicas de respuesta a incidentes incluidas en los contratos con el proveedor, niveles de servicio comprometidos para la colaboración durante un incidente, obligaciones de notificación del proveedor, condiciones de acceso al modelo y a sus componentes por parte de la organización.
Contingencia	Procedimiento alternativo documentado para cada proceso dependiente, personal capacitado para activar y operar el procedimiento alternativo, limitaciones conocidas del procedimiento alternativo respecto a la operación normal con el sistema de IA.

TABLA 10: ELEMENTOS DEL INVENTARIO DE SISTEMAS DE IA

El inventario debe revisarse y actualizarse de forma periódica y, en todo caso, cuando se produzcan cambios relevantes en el sistema, en los procesos dependientes, en los acuerdos contractuales o en el marco regulatorio aplicable.

9.1.2. Plan de respuesta con especificidades de IA

El plan de respuesta a incidentes de ciberseguridad de la organización debe incorporar procedimientos diferenciados que contemplen las particularidades de los incidentes que pueden afectar a los sistemas de IA. Estos procedimientos no sustituyen los procedimientos generales de respuesta, sino que los extienden para cubrir las tipologías de incidentes, los componentes y los impactos específicos descritos en las secciones 5 y 6 de esta guía.

El plan debe contemplar procedimientos específicos para, al menos, las siguientes tipologías de incidentes, que se corresponden con las categorías identificadas en la sección 6:

- Envenenamiento de datos de entrenamiento o de ajuste fino del modelo.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

- Manipulación o sustitución no autorizada de un modelo en producción.
- Ataque adversario sobre las entradas del sistema en fase de inferencia.
- Compromiso de los pipelines de MLOps (entrenamiento, validación, despliegue).
- Inyección de prompts o manipulación de instrucciones en sistemas de IA generativa.
- Extracción o inversión de modelo.
- Compromiso relevante de la cadena de suministro de componentes de IA.
- Fuga de datos personales o información sensible a través de los resultados del modelo.

Para cada una de estas tipologías, el plan debe definir, al menos, los siguientes elementos: los indicadores que pueden alertar sobre el incidente, las acciones iniciales de contención recomendadas, los perfiles profesionales que deben intervenir conforme al modelo organizativo definido en la sección 8, los criterios de escalado que determinan cuándo debe elevarse la gestión del incidente, y las obligaciones de notificación que pueden resultar aplicables en función de las características del incidente.

El plan de respuesta debe ser un documento vivo, sometido a revisión periódica y actualizado cuando se incorporen nuevos sistemas de IA, cuando cambien las condiciones operativas o contractuales, cuando se identifiquen nuevos vectores de ataque relevantes o cuando las lecciones aprendidas de incidentes anteriores o simulacros así lo aconsejen.

9.1.3. Capacidades técnicas de detección y análisis

La detección de incidentes en sistemas de IA no puede basarse exclusivamente en las herramientas de monitorización de seguridad convencionales desplegadas para la protección de la infraestructura y los sistemas de información. Como se ha descrito en la sección 5, la singularidad de los incidentes con impacto algorítmico reside precisamente en que el sistema puede continuar funcionando sin generar alertas de infraestructura mientras sus resultados están comprometidos. La organización debe desplegar, por tanto, capacidades de monitorización específicas que incluyen, como mínimo:

- **Monitorización continua del rendimiento del modelo.** La organización debe establecer un seguimiento continuo o periódico de las métricas de rendimiento del modelo en producción, comparándolas con los baselines documentados en el inventario del sistema. Una degradación significativa e inexplicada de estas métricas puede ser el primer indicador de que el modelo ha sido comprometido.
- **Controles de integridad sobre los modelos desplegados.** La organización debe implementar mecanismos de verificación periódica que permitan confirmar que el modelo que se encuentra efectivamente desplegado en producción se corresponde con la versión autorizada y validada.
- **Monitorización de la distribución de datos y resultados.** La organización debe contar con capacidades para detectar cambios significativos en la distribución de los datos de entrada que recibe el modelo (data drift) y en la distribución de las salidas que genera (prediction drift). Estos cambios pueden indicar tanto una deriva natural del modelo por evolución del contexto operativo como una manipulación deliberada de los datos o del modelo. La capacidad de distinguir entre ambos escenarios, que se aborda con mayor detalle en el apartado 9.2.3, es un elemento fundamental para una detección eficaz.
- **Monitorización de los pipelines de MLOps**. Todas las operaciones realizadas sobre los pipelines de entrenamiento, validación y despliegue de modelos deben quedar registradas y estar sujetas

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

a supervisión. Esto incluye las modificaciones de código, los cambios en configuraciones e hiperparámetros, los accesos a repositorios de modelos y datos, los despliegues realizados y las validaciones ejecutadas. Cualquier operación no asociada a un cambio autorizado y documentado debe investigarse.

- **Monitorización de las APIs de inferencia.** Los patrones de consulta a las APIs a través de las cuales se accede al sistema de IA deben monitorizarse para detectar comportamientos anómalos que puedan indicar intentos de extracción del modelo mediante consultas sistemáticas, ataques adversarios realizados de forma automatizada o intentos de denegación de servicio. Los indicadores relevantes incluyen el volumen de consultas, su frecuencia, su distribución temporal, la diversidad de las entradas y la presencia de patrones repetitivos o sistemáticos.
- **Monitorización específica de sistemas de IA generativa.** En los sistemas de IA generativa desplegados por la organización, debe supervisarse la generación de contenidos para detectar respuestas inesperadas, contenidos fuera del ámbito previsto, indicios de revelación de información del sistema (como prompts internos o configuraciones), respuestas que sugieran la evasión de los controles de seguridad establecidos o indicios de que el sistema está ejecutando instrucciones no previstas procedentes de inyecciones de prompts.
- **Integración en la plataforma de gestión de eventos de seguridad.** Los registros de actividad generados por los sistemas de IA y por las capacidades de monitorización específicas descritas en los puntos anteriores deben integrarse en el SIEM o en la plataforma de gestión de eventos de seguridad de la organización.

9.1.4. Formación y simulacros

La eficacia de la respuesta a un incidente en un sistema de IA depende de que las personas que deben intervenir comprendan las particularidades de estos incidentes y estén preparadas para actuar conforme a los procedimientos establecidos.

- **Formación de los equipos de respuesta:**

Los equipos de ciberseguridad, los equipos de IA y ciencia de datos y las personas responsables funcionales deben recibir formación específica que incluya: los vectores de ataque específicos de los sistemas de IA conforme a la tipología de la sección 6, los procedimientos de análisis de componentes de IA, los procedimientos de contención específicos y las obligaciones de notificación aplicables conforme a la sección 10.

- **Orientación a personas usuarias y personal profesional:**

Las personas que interactúan directamente con los sistemas de IA deben conocer qué tipos de anomalías deben reportar y a través de qué canal. En entornos de salud, esto incluye la identificación de resultados clínicamente incoherentes o cambios inexplicados en el patrón de recomendaciones. En entornos de Smart Cities, incluye la detección de comportamientos anómalos del sistema o resultados incorrectos.

- **Simulacros:**

La organización debe realizar simulacros periódicos que incluyan escenarios de incidente en sistemas de IA, involucrando a todos los perfiles con responsabilidades en la respuesta. Los escenarios deben incluir, al menos, una situación de incidente con impacto algorítmico no inmediatamente visible, de

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

forma que se ejercite la detección a partir de indicadores sutiles y la toma de decisiones bajo incertidumbre. Los resultados deben documentarse y alimentar la mejora del plan de respuesta.

9.1.5. Acuerdos con proveedores

Cuando el sistema de IA haya sido desarrollado, suministrado o sea mantenido por un proveedor externo, la capacidad de respuesta de la organización ante un incidente depende, en medida significativa, de la colaboración de dicho proveedor. Esta dependencia es particularmente relevante cuando el modelo es propietario del proveedor y la organización no dispone de acceso directo a sus componentes internos ni de capacidad de análisis autónomo sobre el modelo.

Los acuerdos contractuales con proveedores de sistemas de IA deben contemplar las siguientes materias en relación con la gestión de incidentes:

- **Obligaciones de notificación.** El proveedor debe comprometerse a notificar a la organización, de forma inmediata y con la información disponible, cualquier incidente de seguridad, vulnerabilidad o compromiso que afecte al sistema suministrado o a cualquier componente de su cadena de suministro que pueda repercutir sobre la seguridad o la fiabilidad del sistema.
- **Colaboración técnica durante la respuesta.** El contrato debe definir el compromiso del proveedor de proporcionar soporte técnico especializado durante la gestión del incidente, incluyendo niveles de servicio concretos para la disponibilidad de este soporte (tiempo de respuesta, disponibilidad horaria, perfiles técnicos).
- **Preservación y entrega de evidencias.** El proveedor debe comprometerse a preservar y, cuando sea necesario, entregar a la organización los registros, logs, versiones de modelos, datos de entrenamiento y demás evidencias que resulten necesarias para la investigación del incidente.
- **Clarificación de responsabilidades.** El contrato debe establecer con claridad la distribución de responsabilidades entre proveedor y organización en la gestión de incidentes, en coherencia con las obligaciones diferenciadas que el RIA establece para proveedores de sistemas de IA y para las personas responsables de despliegue.
- **Acceso al modelo y a sus componentes**. Cuando el modelo sea propietario del proveedor, los acuerdos deben definir con precisión qué información y qué capacidades de análisis proporcionará el proveedor a la organización durante un incidente, qué análisis realizará el proveedor por su cuenta y en qué plazos comunicará los resultados a la organización. La organización debe evaluar, con carácter previo a la contratación, si las condiciones de acceso ofrecidas por el proveedor resultan suficientes para garantizar una respuesta eficaz a incidentes.

9.2. Detección e identificación

La detección constituye la primera fase activa de la respuesta a un incidente. Su objetivo es identificar, lo antes posible, que un sistema de IA puede estar comprometido o que se ha producido un evento de seguridad que afecta a alguno de sus componentes. La rapidez y la eficacia de la detección condicionan directamente el impacto del incidente.

La detección de incidentes en sistemas de IA presenta una dificultad específica ya que los incidentes con impacto algorítmico se caracterizan precisamente por la posibilidad de que el sistema continúe

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

funcionando con normalidad aparente mientras la fiabilidad de sus resultados ha sido comprometida. En este caso, el sistema no se detiene, no genera necesariamente alertas de infraestructura y puede no mostrar características visibles de fallo desde la perspectiva de la monitorización convencional.

9.2.1. Fuentes de detección

Para ello, la organización debe establecer un modelo de detección que contemple tres categorías de fuentes complementarias:

I. Monitorización técnica automatizada:

Comprende las alertas generadas por los sistemas de monitorización descritos en el apartado 9.1.3. Incluye las alertas procedentes de la infraestructura convencional (accesos no autorizados, tráfico anómalo, modificaciones no previstas en configuraciones), pero también, y de forma especialmente relevante, las alertas generadas por las capacidades de monitorización específicas de los sistemas de IA: degradación del rendimiento del modelo respecto a los baselines documentados, discrepancias en los controles de integridad del modelo desplegado, cambios significativos en la distribución de datos de entrada o de salidas, operaciones no autorizadas en los pipelines de MLOps, patrones anómalos de consulta a las APIs de inferencia y, en sistemas de IA generativa, generación de contenidos que sugieran la evasión de controles o la ejecución de instrucciones no previstas.

II. Personas usuarias y personal profesional:

Las personas que interactúan directamente con los sistemas de IA en su actividad profesional constituyen una fuente de detección que resulta crítica precisamente para aquellos incidentes que son más difíciles de detectar por medios automatizados: los que producen alteraciones en los resultados del modelo que son funcional o clínicamente significativas pero que pueden no reflejarse de forma inmediata en las métricas técnicas de rendimiento.

Para que esta fuente de detección sea efectiva, es necesario que estas personas hayan recibido la orientación prevista en el apartado 9.1.4 sobre qué tipos de anomalías deben reportar, y que la organización haya establecido canales de comunicación claros, accesibles y conocidos a través de los cuales puedan reportar estos indicios de forma directa y ágil al equipo de respuesta.

III. Fuentes externas:

La detección puede originarse también a partir de información procedente del exterior de la organización. Las fuentes externas relevantes incluyen las notificaciones de proveedores sobre vulnerabilidades o incidentes que afecten al sistema suministrado o a componentes de su cadena de suministro, las alertas y avisos de organismos competentes (CSIRT nacionales, ENISA, autoridades de ciberseguridad), la publicación de vulnerabilidades en componentes de código abierto o de terceros utilizados por el sistema, la información de inteligencia de amenazas que identifique campañas o vectores de ataque relevantes para los sistemas de IA desplegados por la organización, y las comunicaciones de otras organizaciones que utilicen sistemas o componentes similares y hayan detectado incidentes que puedan afectar a los mismos componentes. La organización debe mantener actualizados los canales de recepción de esta información y disponer de procedimientos para su evaluación y, cuando proceda, para la activación de la respuesta.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

9.2.2. Indicadores de compromiso específicos de sistemas de IA

Además de los indicadores de compromiso convencionales que los equipos de ciberseguridad monitorizan habitualmente (accesos no autorizados, tráfico de red anómalo, modificaciones no autorizadas de configuraciones, presencia de malware), la detección de incidentes en sistemas de IA requiere la vigilancia de un conjunto de indicadores específicos que se derivan de la naturaleza de estos sistemas y de sus componentes. Estos indicadores están relacionados con las categorías de incidentes descritas en la sección 6 y con las particularidades de los sistemas de IA expuestas en la sección 5.

La siguiente tabla recoge los principales indicadores de compromiso específicos de los sistemas de IA, organizados por categorías:

Categoría	Indicadores
Integridad del modelo	Cambios no autorizados en la versión del modelo desplegado en producción. Discrepancias entre el hash o la firma del artefacto desplegado y el hash o la firma de la versión autorizada. Modificaciones no registradas en los parámetros, los pesos o la configuración del modelo. Presencia de componentes o capas no esperados en la arquitectura del modelo.
Integridad de los datos	Modificaciones no autorizadas en los datos de entrenamiento, validación o ajuste fino. Adición de registros no explicados en los conjuntos de datos. Alteraciones en los datos de entrada que recibe el modelo durante la inferencia que no se corresponden con los patrones normales de operación.
Comportamiento del modelo	Degradación significativa e inexplicada del rendimiento del modelo respecto a los baselines documentados. Cambios estadísticamente significativos en la distribución de las salidas del modelo que no se corresponden con cambios conocidos en los datos de entrada. Discrepancias persistentes entre las predicciones del modelo y los resultados reales observados por personal profesional. Aparición de patrones de error nuevos o concentrados en subconjuntos específicos de datos.
Pipeline de MLOps	Modificaciones en las configuraciones del pipeline no asociadas a un cambio autorizado y documentado. Despliegues de modelos no registrados en el sistema de control de versiones. Alteraciones en las dependencias o librerías utilizadas por el pipeline. Ejecución de procesos de entrenamiento o validación no planificados.
APIs de inferencia	Incremento anómalo del volumen de consultas. Patrones de consulta sistemáticos o automatizados que sugieran intentos de extracción del modelo. Consultas con entradas diseñadas para explorar los límites del modelo o para provocar respuestas específicas. Accesos desde orígenes no autorizados.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

IA generativa	Generación de contenidos inesperados, fuera del ámbito previsto o que revelen información interna del sistema, como prompts del sistema, instrucciones de configuración o datos de entrenamiento. Respuestas que sugieran la evasión de los controles de seguridad establecidos (filtros, guardrails, restricciones temáticas). Indicios de que el sistema está ejecutando instrucciones que no forman parte de su configuración autorizada, lo que puede indicar una inyección de prompts exitosa.
Registros de actividad	Alteraciones, truncamientos, eliminaciones o lagunas inexplicadas en los registros de actividad del sistema de IA o de sus componentes. Inconsistencias entre los registros de diferentes componentes del sistema.

TABLA 11: INDICADORES DE COMPROMISO ESPECÍFICOS DE SISTEMAS DE IA

Es importante destacar que la detección eficaz de incidentes en sistemas de IA rara vez se basa en un único indicador aislado. La efectividad de la detección reside en la capacidad de correlacionar múltiples señales procedentes de fuentes y categorías distintas.

9.2.3. Distinción entre degradación natural y compromiso malicioso

Los sistemas de IA pueden experimentar una degradación progresiva de su rendimiento por causas no maliciosas, principalmente la deriva del modelo (data drift o concept drift). Esta degradación puede producir síntomas similares a los de un compromiso deliberado. Los elementos que pueden ayudar a diferenciar ambas situaciones incluyen:

- **La gradualidad del cambio**. Una degradación progresiva y uniforme es más compatible con una deriva natural. Un cambio abrupto sin causa conocida sugiere un compromiso deliberado.
- **La selectividad de la alteración.** Una degradación uniforme sugiere deriva general. Una alteración concentrada en subconjuntos específicos puede indicar manipulación dirigida.
- **La correlación con cambios conocidos en el entorno.** Si la degradación es explicable por cambios documentados en el contexto operativo, la hipótesis de deriva es más plausible.
- **La presencia de otros indicadores de compromiso.** La existencia simultánea de indicadores de otras categorías refuerza significativamente la hipótesis de compromiso malicioso.
- **La coherencia temporal.** La coincidencia con otros eventos de seguridad, alertas de proveedores o publicación de vulnerabilidades refuerza la hipótesis de compromiso.

La organización no debe exigir certeza absoluta antes de tratar un evento como un incidente de seguridad. Cuando no sea posible descartar un compromiso malicioso con un grado razonable de confianza a partir de los elementos disponibles, el evento debe tratarse como un incidente de seguridad y gestionarse conforme a las fases posteriores de este proceso. La confirmación o descarte definitivo se producirá durante la fase de análisis e investigación (apartado 9.5).

9.3. Clasificación y priorización

Una vez que se ha detectado un evento que confirma o hace razonablemente sospechar la existencia de un incidente de ciberseguridad que afecta a un sistema de IA, debe procederse a su clasificación y

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

priorización. El objetivo de esta fase es determinar la naturaleza del incidente, evaluar su severidad y establecer la intensidad y urgencia de la respuesta que requiere.

La clasificación inicial debe realizarse de forma conjunta por el equipo de ciberseguridad (CSIRT/SOC) y el equipo de IA y ciencia de datos, ya que la determinación del componente afectado, la tipología del incidente y la evaluación del posible impacto algorítmico requieren capacidades que residen en ambos equipos. Cuando los indicios sugieran que el incidente puede afectar a datos personales, el Delegado de Protección de Datos debe incorporarse al proceso de clasificación desde esta fase. Cuando los indicios sugieran una posible afectación a procesos clínicos o a servicios públicos esenciales, la persona responsable funcional del proceso afectado debe ser consultada para valorar la criticidad operativa del incidente.

Es importante señalar que la clasificación inicial es necesariamente provisional, ya que se realiza con la información disponible en las primeras fases de la respuesta, que es por definición incompleta. La clasificación debe revisarse y, en su caso, modificarse a medida que el análisis del incidente aporte información adicional.

9.3.1. Criterios de clasificación

La clasificación del incidente debe realizarse atendiendo a un conjunto de criterios que permitan caracterizar tanto su dimensión técnica como su impacto operativo y sobre las personas. Los criterios que deben evaluarse son, al menos, los siguientes:

Criterio	Elementos para evaluar
Dimensión de seguridad afectada	Confidencialidad de los datos o del modelo, integridad de los datos o del modelo, disponibilidad del sistema, y fiabilidad de los resultados generados por el sistema de IA. Esta última dimensión, específica de los sistemas de IA, debe evaluarse siempre que el incidente pueda haber afectado al comportamiento del modelo o a la calidad de sus salidas.
Vector de ataque	Si el ataque utiliza un vector convencional (explotación de vulnerabilidades de infraestructura, compromiso de credenciales), un vector específico de IA (envenenamiento, ataque adversario, inyección de prompts) o una combinación de ambos.
Componente afectado	Qué componente o componentes del sistema de IA se han visto afectados: datos de entrenamiento, modelo, pipeline de MLOps, API de inferencia, prompts o configuraciones del sistema, resultados generados.
Tipología del incidente	Categoría del incidente conforme a la clasificación establecida en la sección 6 de esta guía.
Impacto sobre las personas	Si el incidente puede afectar a la seguridad de la persona paciente, a la seguridad pública, al ejercicio de derechos fundamentales o a datos personales de categoría especial.

Impacto sobre los servicios	Si el incidente afecta a la continuidad de procesos asistenciales, de servicios urbanos esenciales o de otros servicios públicos.
Alcance	Número de sistemas de IA afectados o potencialmente afectados, número de personas potencialmente afectadas (personas pacientes, ciudadanía, personal profesional) y potencial de propagación del compromiso a otros sistemas o procesos.
Persistencia y reversibilidad	Si el compromiso es persistente (como en el caso de un modelo envenenado que permanece comprometido tras cerrar la vulnerabilidad), si se han adoptado decisiones sobre resultados comprometidos y si dichas decisiones y sus consecuencias son reversibles.

TABLA 12: CRITERIOS DE CLASIFICACIÓN DE INCIDENTES

9.3.2. Niveles de severidad

A partir de la evaluación de los criterios anteriores, el incidente debe asignarse a uno de los siguientes niveles de severidad, que determinan la intensidad de la respuesta, los recursos movilizados y la urgencia de las actuaciones:

Severidad	Criterios de asignación
Crítica	Se asigna cuando concurre alguna de las siguientes circunstancias: – Existe un riesgo inmediato y grave para la vida o la integridad física de las personas como consecuencia del incidente – Se ha producido una afectación grave de un servicio esencial con impacto directo sobre la seguridad pública – Se ha producido una exposición masiva de datos de categoría especial – Se ha confirmado el compromiso de un sistema de IA clasificado como de alto riesgo conforme al RIA con impacto activo sobre decisiones clínicas o de seguridad pública.
Alta	Se asigna cuando: – El incidente compromete la fiabilidad de un sistema de IA de alto riesgo sin que exista un riesgo inmediato para la vida – Cuando produce una afectación significativa de la continuidad de servicios asistenciales o urbanos esenciales – Cuando implica una violación de seguridad de datos personales que afecta a un número significativo de personas o a datos de categoría especial.

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia
Junta de Andalucía
Agencia Digital de Andalucía

Media	Se asigna cuando el incidente afecta al rendimiento o a la seguridad de un sistema de IA sin que se haya producido un impacto inmediato sobre las personas, pero con potencial de agravamiento si no se gestiona adecuadamente. Incluye el compromiso de componentes del pipeline de MLOps sin evidencia de afectación al modelo en producción, los intentos de extracción o inversión de modelo sin confirmación de éxito, o la detección de anomalías en los patrones de consulta a APIs que requieren investigación.
Baja	Se asigna cuando el impacto del incidente es limitado y está controlado. Incluye anomalías detectadas y contenidas sin evidencia de afectación material sobre los resultados del sistema, sobre las personas o sobre los servicios, así como eventos que requieren investigación, pero cuya evaluación inicial no indica un riesgo significativo.

TABLA 13: NIVELES DE SEVERIDAD

La clasificación inicial es provisional y debe revisarse a medida que el análisis del incidente aporte información adicional. Un incidente clasificado inicialmente como de severidad media puede reclasificarse como alta o crítica si el análisis revela un período de compromiso prolongado o un impacto sobre decisiones que no era evidente inicialmente.

9.3.3. Factores específicos de evaluación de impacto en IA

La evaluación del impacto de un incidente en un sistema de IA debe tener en cuenta un conjunto de factores que no están presentes, o no tienen la misma relevancia, en los incidentes que afectan a sistemas de información convencionales. Estos factores son determinantes para una clasificación adecuada del incidente y para la planificación de las fases posteriores de la respuesta.

I. **Persistencia del compromiso.**

En un sistema de información convencional, la eliminación del vector de ataque (cierre de la vulnerabilidad, eliminación del malware, revocación de credenciales comprometidas) suele ser suficiente para resolver el compromiso del sistema. En un sistema de IA, esto no es necesariamente así. Un modelo que ha sido entrenado con datos envenenados o que ha sido manipulado directamente permanece comprometido de forma persistente incluso después de que se haya eliminado la vulnerabilidad que permitió la manipulación. La erradicación del compromiso en el modelo puede requerir el reentrenamiento completo del mismo. Este factor debe considerarse en la evaluación de la severidad y en la planificación de la erradicación.

II. **Período de exposición y acumulación de daño**

Los incidentes con impacto algorítmico, como se ha descrito en la sección 5, pueden haber permanecido activos durante períodos prolongados antes de su detección, durante los cuales el sistema ha continuado generando resultados comprometidos que han alimentado procesos de decisión clínica, de gestión urbana o de prestación de servicios públicos. Cuanto más prolongado

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

sea el período de compromiso, mayor será el volumen de resultados potencialmente alterados y, en consecuencia, mayor será el número de decisiones que pueden haberse visto afectadas. La evaluación del impacto no puede limitarse al momento de la detección, sino que debe considerar retrospectivamente todo el período estimado de compromiso.

III. Reversibilidad de las consecuencias

Las decisiones adoptadas sobre la base de resultados comprometidos pueden haber producido efectos que no se revierten automáticamente con la restauración del sistema a un estado íntegro. Un diagnóstico incorrecto generado por un sistema de apoyo a la decisión clínica puede haber dado lugar a la prescripción de un tratamiento inadecuado o a la no prescripción de un tratamiento necesario. Una priorización alterada generada por un sistema de triaje puede haber retrasado la atención a situaciones que requerían intervención urgente. Una asignación incorrecta de recursos generada por un sistema de gestión urbana puede haber afectado a la calidad del servicio público durante un período prolongado. La evaluación del impacto debe considerar estas consecuencias y su grado de reversibilidad.

IV. Efecto de encadenamiento.

En muchos entornos operativos, la salida de un sistema de IA no es consumida únicamente por un proceso de decisión aislado, sino que puede alimentar a otros procesos, sistemas o decisiones de forma directa o indirecta. Un modelo de triaje que genera priorizaciones incorrectas puede afectar a los sistemas de asignación de recursos que consumen sus resultados. Un modelo de predicción de demanda que genera estimaciones alteradas puede afectar a la planificación de servicios que depende de dichas estimaciones. La evaluación del impacto debe considerar esta propagación funcional y extender el análisis a los procesos y sistemas receptores de las salidas del sistema comprometido.

V. Afectación a la confianza.

Un incidente en un sistema de IA, especialmente si trasciende públicamente o si afecta a procesos sensibles como la asistencia sanitaria o la seguridad pública, puede tener un impacto significativo sobre la confianza del personal profesional, las personas usuarias y la ciudadanía tanto en el sistema concreto afectado como, por extensión, en los procesos en los que se integra y en la adopción de tecnologías de IA en el ámbito público. Este impacto sobre la confianza debe considerarse en la evaluación global del incidente, especialmente en entornos de salud donde la aceptación del sistema por parte del personal clínico condiciona directamente su utilidad y donde la pérdida de confianza puede resultar difícil de restaurar.

9.4. Contención

La contención tiene como objetivo limitar el impacto del incidente impidiendo que el sistema comprometido continúe generando daño. En los incidentes que afectan a sistemas de información convencionales, la contención se orienta fundamentalmente a evitar la propagación del compromiso a otros sistemas y a preservar las evidencias necesarias para la investigación. En los incidentes que afectan a sistemas de IA, la contención debe abordar además una dimensión específica y urgente: mientras el

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

sistema continúa operando con resultados comprometidos, cada inferencia que genera puede alimentar una decisión clínica incorrecta, una actuación inadecuada sobre un servicio público o una respuesta errónea a la ciudadanía. La contención en sistemas de IA no solo busca detener la propagación técnica del compromiso, sino también detener la acumulación de daño funcional derivada de la continuidad operativa de un sistema cuya fiabilidad no está garantizada.

9.4.1. Principio rector

Si existe un riesgo razonable de que el sistema comprometido esté generando resultados que pongan en peligro la seguridad de las personas o la integridad de decisiones relevantes, debe procederse a su suspensión inmediata y a la activación de los procedimientos alternativos previstos. Esta decisión no requiere la confirmación completa de la naturaleza del incidente ni la identificación de la causa raíz. La autoridad para adoptar esta decisión es la definida en el apartado 8.3 de esta guía.

9.4.2. Opciones de contención

La contención de un sistema de IA no es necesariamente una decisión binaria entre mantener el sistema plenamente operativo o suspenderlo por completo. En función de la severidad del incidente, del grado de certeza sobre el compromiso, del componente afectado y de la criticidad del proceso que depende del sistema, pueden considerarse diferentes opciones de contención con distinto grado de intensidad. La selección de la opción adecuada debe realizarse por el coordinador del incidente en consulta con el equipo de IA y la persona responsable funcional del proceso afectado, y debe documentarse junto con su justificación.

Las opciones de contención, ordenadas de menor a mayor intensidad, son las siguientes:

Opción	Descripción	Aplicabilidad
Monitorización reforzada	Se mantiene el sistema operativo con supervisión intensificada de sus resultados	Cuando el compromiso no está confirmado y el riesgo es bajo
Modo supervisado	Se mantiene el sistema operativo pero todos sus resultados son revisados por una persona antes de integrarse en el proceso de decisión	Cuando el sistema es crítico para el proceso y la supervisión humana puede mitigar el riesgo
Suspensión parcial	Se suspenden las funcionalidades o los flujos de datos afectados, manteniendo operativas las funcionalidades no comprometidas	Cuando el compromiso está acotado a un componente o funcionalidad específica
Suspensión completa	Se retira el sistema de producción y se activan los procedimientos alternativos	Cuando existe riesgo para la seguridad de las personas, cuando no es posible acotar el compromiso, cuando la severidad es crítica o alta, o cuando las opciones de menor intensidad no son viables.

TABLA 14: OPCIONES DE CONTENCIÓN

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

Cuando no exista un procedimiento alternativo documentado y operativo para el proceso afectado, esta circunstancia debe tenerse en cuenta en la evaluación de la opción de contención, pero en ningún caso debe impedir la suspensión del sistema si existe un riesgo para la seguridad de las personas. En estos casos, la organización debe establecer, de forma inmediata, medidas provisionales que permitan mantener el proceso con las garantías mínimas necesarias, aceptando que estas medidas provisionales pueden implicar una reducción temporal de la capacidad operativa o del nivel de servicio.

9.4.3. Acciones de contención sobre la infraestructura y los accesos

Con independencia de la opción de contención seleccionada sobre el sistema de IA, deben adoptarse las acciones de contención necesarias sobre la infraestructura y los accesos para detener la progresión del compromiso técnico. Estas acciones incluyen, según las circunstancias del incidente:

- El aislamiento de los entornos que alojan el sistema de IA, para evitar la propagación del compromiso a otros sistemas o entornos de la organización. El alcance del aislamiento debe ser proporcionado al alcance conocido o sospechado del compromiso, y debe tener en cuenta las dependencias del sistema con otros componentes de la infraestructura.
- El bloqueo o la revocación de las credenciales comprometidas o sospechosas de compromiso, incluyendo credenciales de acceso a los entornos que alojan el sistema, a los repositorios de modelos y datos, a las APIs de inferencia y a los pipelines de MLOps.
- La restricción del acceso a las APIs de inferencia del sistema, para evitar que el sistema continúe siendo consultado mientras se evalúa el compromiso, o para impedir que un atacante continúe explotando las APIs como vector de ataque.

La preservación del estado de los sistemas afectados antes de introducir cualquier cambio que pueda destruir o alterar evidencias relevantes para la investigación posterior. Esta preservación debe realizarse conforme a los procedimientos de recogida de evidencias detallados en el apartado 9.4.7.

9.4.4. Acciones de contención sobre los componentes de IA

Además de las acciones sobre la infraestructura, la contención debe incluir acciones específicas sobre los componentes propios del sistema de IA:

- La suspensión o restricción del sistema de IA conforme a la opción de contención seleccionada en el apartado 9.4.2, asegurando que la retirada del sistema se produce de forma controlada y documentada.
- La retirada de los resultados del sistema de los flujos de decisión afectados. No basta con suspender el sistema; debe verificarse que los procesos que consumían sus resultados han dejado efectivamente de hacerlo y que no existen resultados pendientes de procesamiento que puedan ser utilizados inadvertidamente tras la suspensión.
- La congelación de los pipelines de reentrenamiento, actualización y despliegue de modelos. Esta acción es fundamental para evitar que un compromiso activo en los datos de entrenamiento o en el pipeline produzca nuevas versiones del modelo que incorporen el compromiso, lo que podría perpetuar el problema incluso después de la restauración del modelo en producción.
- El bloqueo del acceso a los repositorios de modelos y de datos de entrenamiento, para evitar que el atacante pueda continuar modificando estos componentes o que se realicen operaciones no controladas sobre los mismos durante la gestión del incidente.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

- La preservación del estado completo del modelo en producción en el momento del incidente, incluyendo los pesos del modelo, su arquitectura, su configuración, sus metadatos y la versión exacta del artefacto desplegado.
- La preservación de los datos de inferencia correspondientes al período sospechoso, incluyendo las entradas recibidas por el modelo, las salidas generadas y las marcas temporales asociadas. Estos datos resultan esenciales para la determinación posterior del alcance temporal del compromiso y para la evaluación del impacto sobre las decisiones adoptadas.
- En sistemas de IA generativa, la preservación adicional de los prompts del sistema, las plantillas de instrucciones, las configuraciones de seguridad (filtros, guardrails, restricciones) y, cuando sea posible, los logs de las conversaciones o interacciones relevantes durante el período sospechoso.

9.4.5. Acciones de contención sobre los procesos dependientes

La contención no se limita al ámbito técnico. Debe extenderse a los procesos operativos, asistenciales o de servicio público que dependen de los resultados del sistema de IA comprometido, para asegurar que estos procesos pueden continuar funcionando con las garantías necesarias mientras el sistema permanece suspendido o restringido.

En entornos de salud:

Activación del procedimiento clínico alternativo, comunicación inmediata e inequívoca al personal sanitario de que el sistema no debe utilizarse o de que sus resultados están sujetos a revisión obligatoria, e identificación preliminar de las decisiones clínicas potencialmente afectadas durante el período de sospecha.

En entornos de Smart Cities:

Activación del procedimiento operativo alternativo, evaluación del impacto inmediato sobre el servicio público y adopción de medidas provisionales para mantener la continuidad del servicio, coordinando con las personas responsables funcionales la transición al modo alternativo.

9.4.6. Notificación al proveedor

Cuando el sistema de IA afectado haya sido desarrollado, suministrado o sea mantenido por un proveedor externo, debe notificarse al proveedor de forma inmediata conforme a los acuerdos contractuales previstos en el apartado 9.1.5 de esta guía. Esta notificación tiene una doble finalidad.

Se debería seguir los siguientes pasos:

I. **Colaboración técnica del proveedor en la gestión del incidente:** el proveedor puede disponer de información, capacidades de análisis y conocimiento del sistema que resulten necesarios para la investigación del incidente y para la erradicación del compromiso, especialmente cuando el modelo es propietario y la organización no tiene acceso directo a sus componentes internos.

II. **Informar al proveedor del posible compromiso de un sistema bajo su responsabilidad técnica o contractual:** Cuando el compromiso pueda tener su origen en componentes proporcionados por el proveedor o en su cadena de suministro, la notificación resulta

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

especialmente relevante para que el proveedor pueda evaluar si otros sistemas, otros clientes o otros componentes de su cadena de suministro pueden estar afectados por el mismo compromiso.

La notificación debe documentarse, incluyendo la fecha, la hora, el medio utilizado, la persona contactada y la información proporcionada.

9.4.7. Preservación de evidencias

Las acciones de contención deben ejecutarse de forma que se preserve la integridad de las evidencias necesarias para las fases posteriores de análisis, investigación y, en su caso, para eventuales procedimientos legales o administrativos. La preservación de evidencias no es una actividad secundaria que pueda posponerse: las primeras acciones de contención pueden destruir o alterar evidencias si no se ejecutan con la debida precaución.

Antes de realizar cualquier modificación sobre los componentes del sistema, deben preservarse, con cadena de custodia documentada, al menos las siguientes evidencias:

Categoría	Evidencias a preservar
Modelo	Copia íntegra del modelo en producción en el momento del incidente, incluyendo pesos, arquitectura, configuración completa y metadatos de versión del artefacto desplegado.
Datos	Datos de entrenamiento, validación y ajuste fino en su estado disponible. Cuando el volumen haga inviable una copia completa inmediata, debe preservarse como mínimo la información necesaria para verificar la integridad del conjunto (hashes, metadatos, registros de modificaciones).
Datos de inferencia	Entradas recibidas por el modelo, salidas generadas y marcas temporales correspondientes al período sospechoso.
Pipelines de MLOps	Logs completos de los pipelines correspondientes al período relevante, incluyendo registros de entrenamiento, validación, despliegue, accesos y modificaciones.
APIs de inferencia	Logs de las APIs, incluyendo consultas recibidas, respuestas generadas, marcas temporales y metadatos asociados.
Métricas	Métricas históricas de rendimiento del modelo, incluyendo los valores monitorizados durante el período anterior y posterior a la detección.
Accesos	Registros de acceso a todos los componentes críticos del sistema: repositorios de modelos, repositorios de datos, pipelines, APIs y entornos de ejecución.
Configuraciones	Configuraciones de preprocesamiento y postprocesamiento de datos vigentes en el momento del incidente.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

IA generativa	Prompts del sistema, plantillas de instrucciones, configuraciones de seguridad (filtros, guardrails, restricciones) y, cuando sea posible, logs de conversaciones o interacciones relevantes durante el período sospechoso.
Infraestructura	Evidencias forenses convencionales de los sistemas técnicos afectados: imágenes de disco, volcados de memoria y logs de red.

TABLA 15: EVIDENCIAS ESPECÍFICAS A PRESERVAR EN INCIDENTES EN SISTEMAS DE IA

Todas las evidencias deben preservarse con una cadena de custodia documentada que registre quién ha recogido cada evidencia, cuándo, en qué condiciones, dónde se almacena y quién tiene acceso a la misma.

9.5. Análisis e investigación

La fase de análisis e investigación tiene como objetivo determinar con el mayor grado de detalle posible la causa raíz del incidente, el alcance completo del compromiso, el período durante el cual el sistema estuvo afectado, los componentes comprometidos y el impacto real sobre los resultados del sistema, sobre las decisiones adoptadas a partir de dichos resultados y sobre las personas y los servicios afectados. Esta fase requiere la colaboración entre los equipos de ciberseguridad y los equipos de IA y ciencia de datos, tal como están definidos en la sección 8.

Cuando el modelo sea propietario del proveedor y la organización no disponga de acceso directo al modelo ni de capacidad de análisis autónomo sobre sus componentes internos, el análisis de los componentes de IA dependerá en gran medida de la colaboración del proveedor conforme a los acuerdos contractuales establecidos en el apartado 9.1.5. En estos casos, la organización debe exigir al proveedor la información y los resultados de análisis necesarios para evaluar el impacto del incidente sobre sus procesos y servicios, y debe documentar tanto las solicitudes realizadas como la información recibida. La dependencia del proveedor para el análisis no exime a la organización de su responsabilidad de evaluar el impacto del incidente sobre las personas y los servicios bajo su competencia.

9.5.1. Análisis técnico de los entornos de soporte

El análisis forense de los entornos técnicos que soportan el sistema de IA constituye habitualmente el punto de partida de la investigación. Su objetivo es determinar el vector de compromiso inicial, reconstruir la secuencia de acciones del atacante y establecer el alcance del compromiso sobre la infraestructura, como paso previo y complementario al análisis de los componentes específicos de IA.

Este análisis incluye, según las circunstancias del incidente:

- La revisión de los logs de acceso, autenticación y actividad en los sistemas que alojan el modelo, los datos de entrenamiento e inferencia y los pipelines de MLOps. El análisis debe buscar accesos no autorizados, escaladas de privilegios, accesos desde orígenes inusuales y actividades realizadas con credenciales comprometidas.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

- El análisis del tráfico de red relevante, orientado a identificar comunicaciones con servidores de comando y control, exfiltraciones de datos o modelos, o tráfico asociado a la explotación de vulnerabilidades.
- La reconstrucción de la cronología de eventos para establecer la secuencia completa del compromiso: cuándo se produjo el acceso inicial, qué acciones se realizaron, en qué orden, durante cuánto tiempo y qué componentes del sistema fueron accedidos o modificados.
- La identificación, cuando proceda, de malware, herramientas de ataque, explotación de vulnerabilidades conocidas o técnicas de persistencia utilizadas por el atacante.
- La determinación de la relación entre el compromiso del entorno técnico y la afectación a los componentes específicos de IA. Es fundamental establecer si el compromiso de la infraestructura ha sido el medio utilizado para acceder a los componentes de IA (por ejemplo, el compromiso de un servidor para acceder al repositorio de modelos) o si, por el contrario, los componentes de IA han sido el objetivo directo del ataque (por ejemplo, un ataque adversario dirigido directamente contra la API de inferencia).

9.5.2. Análisis de los componentes de IA

El análisis de los componentes de IA constituye la aportación más específica y diferencial de esta guía respecto de los procedimientos convencionales de gestión de incidentes. Su objetivo es determinar si el modelo, los datos o los pipelines han sido comprometidos, de qué forma y con qué consecuencias sobre los resultados generados por el sistema.

I. **Análisis del modelo:**
   - **Verificación de integridad:** comparación del modelo en producción con la última versión validada y autorizada, mediante la verificación de hashes criptográficos, firmas digitales o metadatos de versión del artefacto. Cualquier discrepancia confirma que el modelo ha sido modificado sin autorización.
   - **Evaluación del rendimiento:** ejecución del modelo sobre los conjuntos de datos de referencia (test sets) utilizados durante la validación original, y comparación de los resultados obtenidos con los baselines documentados en el inventario del sistema. Esta evaluación debe analizar tanto las métricas globales de rendimiento (precisión, sensibilidad, especificidad, calibración u otras según la naturaleza del modelo) como el rendimiento desagregado por categorías o subgrupos relevantes, ya que un ataque dirigido puede alterar el comportamiento del modelo únicamente para determinados tipos de entrada sin afectar significativamente a las métricas globales.
   - **Detección de puertas traseras:** análisis del comportamiento del modelo ante entradas específicamente diseñadas para activar posibles puertas traseras (triggers) que puedan haber sido introducidas mediante ataques de envenenamiento. Este análisis puede incluir técnicas de interpretabilidad del modelo para identificar patrones de activación anómalos, la evaluación del comportamiento del modelo ante variaciones controladas de las entradas o la comparación del comportamiento del modelo comprometido con el de un modelo de referencia limpio.
   - **Evaluación de robustez adversaria:** cuando el tipo de incidente lo sugiera, evaluación de la sensibilidad del modelo ante perturbaciones adversarias en las entradas, para determinar si el modelo presenta vulnerabilidades explotables mediante ataques de evasión.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

II. **Análisis de los datos:**

- **Verificación de integridad de los datos de entrenamiento, validación y ajuste fino:** detección de adiciones, modificaciones o eliminaciones no autorizadas en los conjuntos de datos. Este análisis puede apoyarse en la comparación de hashes de los conjuntos de datos, en la revisión de los registros de modificación y en el análisis estadístico de la distribución de los datos para detectar anomalías.
- **Análisis de los datos de inferencia del período sospechoso:** examen de las entradas recibidas por el modelo durante el período de compromiso para identificar posibles entradas adversarias diseñadas para provocar resultados incorrectos o para explotar vulnerabilidades del modelo.
- **Evaluación de la consistencia de las salidas:** comparación de las salidas generadas por el modelo durante el período sospechoso con las salidas que cabría esperar para las entradas recibidas, con el fin de determinar el volumen y la naturaleza de los resultados potencialmente alterados.

III. **Análisis de los pipelines de MLOps:**

- Revisión de todos los accesos y modificaciones realizados sobre el pipeline durante el período relevante, incluyendo cambios en el código, en las configuraciones, en los hiperparámetros y en las automatizaciones.
- Verificación de la integridad de las dependencias, librerías y componentes de terceros utilizados en el pipeline. El compromiso de una dependencia puede constituir un vector de ataque a través de la cadena de suministro que resulte difícil de detectar mediante el análisis de los componentes principales del pipeline.
- Comprobación de que los despliegues de modelos realizados durante el período sospechoso corresponden a versiones autorizadas y validadas, y de que los procesos de validación previos al despliegue se ejecutaron correctamente y no fueron alterados o eludidos.
- Análisis específico de sistemas de IA generativa. Cuando el incidente afecte a un sistema de IA generativa, el análisis debe incluir, además de los elementos anteriores que resulten aplicables:
    - Revisión de los prompts del sistema y las plantillas de instrucciones para detectar modificaciones no autorizadas, adiciones de instrucciones maliciosas o alteraciones que puedan haber facilitado la evasión de los controles de seguridad.
    - Análisis de las respuestas generadas por el sistema durante el período sospechoso para identificar patrones de evasión de controles, generación de contenido no autorizado, revelación de información confidencial del sistema (como prompts internos, datos de entrenamiento o configuraciones) o ejecución de instrucciones no previstas procedentes de inyecciones de prompts
    - Evaluación de la integridad y la eficacia de las configuraciones de seguridad del sistema (filtros de contenido, guardrails, restricciones temáticas, mecanismos de detección de inyecciones) para determinar si han sido alteradas o si presentan debilidades que hayan sido explotadas.
    - Cuando el sistema utilice un modelo fundacional proporcionado por un tercero, determinación de si el compromiso puede tener su origen en el modelo fundacional

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

subyacente y, en su caso, coordinación con el proveedor del modelo para su evaluación conforme a los acuerdos contractuales establecidos.

9.5.3. Determinación del alcance temporal y funcional

La determinación del período durante el cual el sistema estuvo comprometido y del alcance funcional del compromiso es una de las tareas más relevantes y, a menudo, más complejas de la fase de análisis. Sus resultados condicionan directamente el alcance de la revisión de decisiones que deberá realizarse en la fase de recuperación (apartado 9.7.3), el contenido de las notificaciones a autoridades y personas afectadas (sección 10) y la evaluación global del impacto del incidente.

El equipo de análisis debe determinar, como mínimo, los siguientes elementos:

- **Fecha más temprana con indicios de compromiso**. Esta determinación debe realizarse mediante la correlación de los indicadores técnicos identificados en el análisis forense, las anomalías detectadas en el comportamiento del modelo, los registros de acceso y modificación de componentes y cualquier otra evidencia disponible.
- **Fecha de detección efectiva.** La fecha y hora en que el incidente fue detectado o en que se recibió la primera alerta o comunicación que dio lugar a la investigación.
- **Volumen de inferencias realizadas durante el período de compromiso.** El número de consultas procesadas por el sistema, de resultados generados y de resultados que fueron efectivamente integrados en procesos de decisión durante el período comprendido entre la fecha estimada de compromiso y la fecha de contención.
- **Procesos de decisión afectados.** La identificación de los procesos concretos que utilizaron los resultados del sistema durante el período de compromiso.
- **Propagación funcional.** La determinación de si los resultados generados por el sistema comprometido han alimentado a otros sistemas, procesos o decisiones de forma directa o indirecta, extendiendo así el alcance del compromiso más allá del sistema inicialmente afectado.
- **Personas afectadas.** La identificación, en la medida de lo posible, de las personas concretas cuyos datos personales han podido verse comprometidos o cuyas decisiones asociadas (clínicas, de servicio público, administrativas) pueden haberse visto afectadas por resultados comprometidos del sistema.

9.5.4. Evaluación del impacto sobre personas y servicios

La evaluación del impacto debe integrar la información técnica obtenida en los apartados anteriores con la valoración funcional del proceso afectado, realizada con la participación de las personas responsables funcionales y, cuando proceda, del personal profesional cualificado en la materia. Esta evaluación no puede ser realizada exclusivamente por los equipos técnicos, ya que la determinación de la relevancia clínica de una alteración en los resultados de un sistema de apoyo al diagnóstico o de la relevancia funcional de una alteración en los resultados de un sistema de gestión urbana requiere conocimientos específicos del ámbito de aplicación.

En entornos de salud:

La evaluación del impacto debe incluir la identificación de las personas pacientes cuyos procesos asistenciales han podido verse afectados por resultados comprometidos del sistema de IA, la evaluación

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

de la relevancia clínica de las posibles alteraciones en los resultados realizada con la participación de personal profesional sanitario cualificado, la determinación de si existe un riesgo que requiera actuación asistencial inmediata sobre alguna persona paciente (situación que debe gestionarse con la máxima prioridad), y la evaluación de si el incidente ha supuesto una violación de seguridad de datos de salud que deba valorarse conforme al RGPD como brecha de datos personales.

En entornos de Smart Cities:

La evaluación del impacto debe incluir la identificación del servicio público afectado y la determinación del impacto real sobre la ciudadanía, la evaluación de si las decisiones o actuaciones adoptadas sobre la base de los resultados comprometidos han generado un perjuicio real o potencial para la ciudadanía o para terceros, la determinación de si el incidente ha podido afectar a datos personales de la ciudadanía y, en su caso, la valoración de la brecha conforme al RGPD, y la evaluación de la necesidad de adoptar medidas correctoras inmediatas sobre el servicio afectado para mitigar los perjuicios producidos.

9.6. Erradicación

La erradicación tiene como objetivo eliminar la causa raíz del incidente y asegurar que todos los componentes afectados del sistema de IA han sido restaurados a un estado íntegro y verificado. En los incidentes que afectan a sistemas de información convencionales, la erradicación se centra habitualmente en la eliminación del malware, el cierre de las vulnerabilidades explotadas y la restauración de las configuraciones comprometidas. En los incidentes que afectan a sistemas de IA, la erradicación presenta una complejidad adicional que debe comprenderse y gestionarse adecuadamente: la eliminación del vector de ataque no garantiza la eliminación del compromiso en el sistema.

La erradicación en sistemas de IA debe abordar de forma diferenciada tanto la infraestructura y los accesos como cada uno de los componentes específicos de IA que hayan resultado afectados.

9.6.1. Erradicación en la infraestructura y los accesos

Las acciones de erradicación sobre la infraestructura y los accesos siguen los procedimientos habituales de gestión de incidentes de ciberseguridad, adaptados a las circunstancias concretas del incidente. Incluyen, según el caso:

- Eliminación del malware o de las herramientas de ataque identificadas durante la fase de análisis, verificando que la eliminación es completa y que no persisten mecanismos de persistencia que permitan al atacante recuperar el acceso.
- Cierre de las vulnerabilidades explotadas, mediante la aplicación de los parches o las correcciones necesarias, o mediante la implementación de controles compensatorios cuando el parcheado inmediato no sea posible.
- Restauración de las configuraciones seguras en los sistemas, servicios y componentes de infraestructura que hayan sido modificados durante el compromiso.
- Rotación de todas las credenciales que hayan podido verse comprometidas, incluyendo las credenciales de acceso a los entornos que alojan el sistema de IA, a los repositorios de modelos y datos, a las APIs de inferencia, a los pipelines de MLOps y a cualquier otro componente al que

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

el atacante haya podido acceder. La rotación debe ser completa y no limitarse a las credenciales de las que se tenga constancia de compromiso, ya que la determinación exhaustiva de todas las credenciales afectadas puede no ser posible con certeza absoluta.

- Saneamiento de los entornos afectados, que puede incluir, en los casos más graves, la reconstrucción completa de los entornos comprometidos a partir de imágenes o configuraciones de referencia verificadas.

9.6.2. Erradicación en los componentes de IA

La erradicación en los componentes de IA debe abordarse de forma diferenciada para cada tipo de componente, ya que las acciones necesarias y su complejidad varían significativamente en función de la naturaleza del compromiso.

I. **Modelo comprometido.** Cuando el análisis haya confirmado o no haya podido descartar que el modelo en producción ha sido comprometido, deben adoptarse las siguientes acciones:
   - Retirada definitiva del modelo comprometido de producción, si no se había retirado ya durante la fase de contención.
   - Restauración de la última versión íntegra del modelo que haya sido verificada como no afectada por el compromiso. Esta verificación debe ser rigurosa: no basta con restaurar una versión anterior sin confirmar que dicha versión es anterior al inicio del compromiso y que no fue generada a partir de datos o pipelines que estuvieran ya comprometidos en el momento de su creación. El equipo de IA debe verificar la integridad del modelo restaurado mediante la comparación con los registros de versiones autorizadas y mediante la evaluación de su rendimiento frente a los baselines documentados.
   - Cuando no sea posible verificar la integridad de ninguna versión anterior del modelo, o cuando el compromiso afecte a los datos de entrenamiento de forma que cualquier modelo entrenado con dichos datos pueda estar afectado, será necesario proceder al reentrenamiento completo del modelo desde datos verificados como íntegros.

II. **Datos de entrenamiento comprometidos.** Cuando el análisis haya determinado que los datos de entrenamiento, validación o ajuste fino del modelo han sido manipulados, las acciones de erradicación deben incluir:
   - Identificación y eliminación de los datos manipulados, inyectados o alterados
   - Verificación de la integridad del conjunto de datos resultante tras la eliminación de los datos comprometidos, para confirmar que el conjunto resultante es coherente, completo en la medida de lo posible y apto para el entrenamiento o la validación del modelo.
   - Nuevo entrenamiento del modelo a partir del conjunto de datos limpio, o como mínimo, validación exhaustiva del modelo existente frente al conjunto de datos verificado para confirmar que su comportamiento no ha sido alterado de forma significativa. La decisión entre reentrenar y validar debe basarse en la evaluación del alcance del compromiso en los datos y del riesgo residual que pueda persistir.
   - Cuando no sea posible determinar con certeza qué datos han sido comprometidos, la organización debe evaluar la necesidad de reconstruir el conjunto de entrenamiento desde sus fuentes originales verificadas.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

III. **Pipelines de MLOps comprometidos.** Cuando el análisis haya determinado que los pipelines de entrenamiento, validación o despliegue han sido comprometidos, las acciones de erradicación deben incluir:

- Revisión completa de los repositorios de código, las configuraciones, los scripts de automatización y los procesos de validación que componen el pipeline, para identificar y eliminar todas las modificaciones no autorizadas.
- Verificación de la integridad de todas las dependencias, librerías y componentes de terceros utilizados por el pipeline, incluyendo la comprobación de las versiones instaladas frente a las versiones autorizadas, la verificación de hashes o firmas de los paquetes y, cuando sea posible, la revisión de los cambios introducidos en las versiones utilizadas. Los modelos preentrenados descargados de repositorios públicos deben someterse al mismo nivel de verificación.
- Restauración del pipeline desde fuentes de confianza verificadas, cuando la revisión y corrección punto por punto no ofrezca garantías suficientes de que todas las modificaciones maliciosas han sido identificadas y eliminadas.
- Refuerzo de los controles de acceso al pipeline, de los mecanismos de verificación de integridad de sus componentes y de los procesos de validación previos al despliegue, para reducir la posibilidad de que un compromiso similar se repita.

IV. **Componentes de IA generativa.** Cuando el incidente haya afectado a un sistema de IA generativa, la erradicación debe incluir, además de las acciones anteriores que resulten aplicables:

- Revisión y restauración de los prompts del sistema, las plantillas de instrucciones y las configuraciones de seguridad (filtros de contenido, guardrails, restricciones temáticas, mecanismos de detección de inyecciones) a su estado autorizado y verificado, comprobando que no persisten modificaciones no autorizadas.
- Verificación de que los filtros, restricciones y guardrails del sistema funcionan correctamente tras la restauración, mediante la ejecución de pruebas específicas.
- Cuando el compromiso afecte al modelo fundacional proporcionado por un tercero, coordinación con el proveedor para la aplicación de las correcciones necesarias, la actualización a una versión no afectada o, en su caso, la sustitución del modelo. La organización debe exigir al proveedor la información necesaria para evaluar si la corrección aplicada es suficiente para resolver el compromiso.

V. **APIs de inferencia.** Cuando las APIs de inferencia del sistema hayan sido explotadas como vector de ataque o se hayan visto comprometidas, las acciones de erradicación deben incluir:

- Rotación de todas las credenciales y tokens de acceso a las APIs.
- Refuerzo de los mecanismos de autenticación y autorización, revisando los permisos otorgados y aplicando el principio de mínimo privilegio.
- Implementación o refuerzo de los mecanismos de limitación de tasa para dificultar los ataques de extracción de modelo y los ataques adversarios automatizados.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

- Refuerzo de la validación de las entradas recibidas por las APIs, para detectar y rechazar entradas que presenten características compatibles con ataques adversarios o con intentos de inyección.

9.6.3. Verificación de la erradicación

Antes de proceder a la fase de recuperación, la organización debe verificar que la erradicación ha sido completa y que todos los componentes restaurados se encuentran en un estado íntegro y seguro. Esta verificación es un requisito previo imprescindible para la reactivación del sistema y no debe omitirse ni realizarse de forma superficial, independientemente de la presión que pueda existir para restaurar la operación del sistema.

La verificación debe incluir, al menos, las siguientes comprobaciones:

- Verificación de la integridad de todos los componentes restaurados del sistema de IA, incluyendo el modelo, los datos, el pipeline, las APIs y, en sistemas de IA generativa, los prompts y las configuraciones de seguridad.
- Validación del rendimiento del modelo restaurado frente a los conjuntos de datos de referencia utilizados durante la validación original, y comparación de los resultados obtenidos con los baselines documentados en el inventario del sistema.
- Ejecución de pruebas específicas orientadas a verificar que las vulnerabilidades o vectores de ataque explotados durante el incidente han sido efectivamente cerrados y que los controles de seguridad reforzados funcionan correctamente.
- Verificación de que las dependencias y componentes de la cadena de suministro utilizados en la restauración son íntegros y corresponden a versiones autorizadas y verificadas.
- En sistemas de IA generativa, ejecución de pruebas con prompts específicamente diseñados para verificar que los controles de seguridad del sistema funcionan correctamente y que las técnicas de inyección o evasión detectadas durante el incidente ya no son efectivas.
- Confirmación formal por parte del equipo de IA y ciencia de datos de que el modelo restaurado produce resultados fiables y de que no se han detectado indicios de compromiso residual.

Los resultados de todas las verificaciones deben documentarse y formar parte del expediente del incidente.

9.7. Recuperación

La recuperación tiene como objetivo restablecer la operación normal del sistema de IA y de los procesos que dependen de él, una vez que la erradicación ha sido completada y verificada. En los incidentes que afectan a sistemas de información convencionales, la recuperación se centra en la restauración de los servicios y en la verificación de su correcto funcionamiento. En los incidentes que afectan a sistemas de IA, la recuperación debe abordar además una dimensión que resulta específica y que puede ser la más relevante en términos de impacto sobre las personas: la revisión de las decisiones que fueron adoptadas sobre la base de los resultados del sistema durante el período en que estuvo comprometido.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

9.7.1. Criterios de reactivación

La reactivación de un sistema de IA tras un incidente es una decisión deliberada que no debe producirse de forma automática una vez completada la erradicación. La presión por restaurar la operación normal del sistema no debe conducir a una reactivación prematura que comprometa la seguridad de las personas o la fiabilidad de los resultados.

La reactivación requiere que se hayan verificado todas las siguientes condiciones:

- La causa raíz del incidente ha sido identificada y erradicada de forma efectiva.
- Todos los componentes del sistema han sido restaurados y su integridad ha sido verificada conforme al apartado 9.6.3.
- El rendimiento del modelo ha sido validado en un entorno de preproducción o equivalente, confirmando que produce resultados fiables y coherentes con los baselines documentados.
- Los controles de seguridad han sido reforzados en las áreas que el análisis del incidente ha identificado como insuficientes, y los refuerzos han sido verificados.
- Las obligaciones de notificación a las autoridades competentes han sido cumplidas o se encuentran en curso de cumplimiento conforme a los plazos aplicables.
- Las personas responsables de los procesos afectados han sido informadas de la inminente reactivación del sistema, de las condiciones bajo las cuales se produce y de las medidas de supervisión reforzada que se aplicarán durante el período de estabilización.
- La reactivación ha sido autorizada formalmente por las personas responsables competentes conforme a la autoridad de reactivación definida en el apartado 8.3.

La verificación de estas condiciones y la decisión de reactivación deben documentarse de forma explícita en el expediente del incidente.

9.7.2. Reintegración progresiva

La reactivación debe realizarse de forma progresiva, con un período de estabilización durante el cual se apliquen medidas de supervisión reforzada. El período debe definirse en función de la severidad del incidente y la criticidad de los procesos dependientes. Las medidas durante este período incluyen:

- **Monitorización reforzada** de métricas de rendimiento, distribución de salidas, integridad de componentes y registros de actividad, con frecuencia significativamente superior a la habitual y umbral de alerta más bajo.
- **Criterios de reversión predefinidos** comunicados a todos los equipos, que establezcan las condiciones bajo las cuales el sistema será suspendido nuevamente.
- **Incremento progresivo del alcance operativo** cuando sea posible, reactivando inicialmente en un subconjunto de procesos o personas usuarias y ampliando conforme se confirme la estabilidad.
- **Reducción gradual de la supervisión humana** conforme se acumule evidencia de correcto funcionamiento, documentando la decisión de pasar a operación normal.

9.7.3. Revisión de decisiones adoptadas durante el período de compromiso

Esta acción constituye una de las diferencias más relevantes y específicas de la respuesta a incidentes en sistemas de IA respecto de la respuesta a incidentes en sistemas de información convencionales. En un incidente que ha afectado a un sistema de IA cuyos resultados han alimentado procesos de decisión,

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

la recuperación debe incluir necesariamente la revisión de las decisiones que fueron adoptadas sobre la base de resultados que pueden haber sido comprometidos.

Esta revisión es una consecuencia directa del principio de que los resultados de un sistema de IA comprometido no deben considerarse fiables, y su necesidad es proporcional al impacto potencial de las decisiones afectadas sobre las personas. Su alcance viene determinado por los resultados de la fase de análisis, en particular por la determinación del período de compromiso, del volumen de inferencias afectadas y de los procesos de decisión que utilizaron dichos resultados (apartado 9.5.3).

En entornos de salud:

Identificación de las personas pacientes cuyas decisiones asistenciales pudieron verse afectadas, cruzando registros de inferencia con registros clínicos. Priorización basada en riesgo clínico. Revisión por personal profesional sanitario cualificado, documentando los hallazgos, la valoración de si el resultado comprometido influyó en la decisión clínica, la evaluación del perjuicio y las acciones correctoras. Cuando se determine perjuicio real o potencial, actuación conforme a los protocolos asistenciales aplicables.

En entornos de Smart Cities:

Identificación de las decisiones, actuaciones o asignaciones realizadas sobre resultados comprometidos. Evaluación del perjuicio para la ciudadanía o la calidad del servicio. Adopción de medidas correctoras proporcionadas. Evaluación de la necesidad de comunicación a afectados identificables o comunicación pública cuando el perjuicio afecte al servicio de forma general.

9.7.4. Comunicación de la reactivación

La reactivación del sistema debe comunicarse de forma explícita a todas las partes implicadas antes de que se produzca efectivamente. Esta comunicación debe incluir:

- **A los equipos internos y personal profesional que interactúan con el sistema.** Debe informarse de que el sistema vuelve a estar operativo, de las condiciones bajo las cuales se produce la reactivación (incluyendo las medidas de monitorización reforzada y, en su caso, el modo de operación supervisado durante el período de estabilización), de cualquier cambio relevante en el funcionamiento del sistema respecto a la situación anterior al incidente y de los criterios de reversión establecidos para el período de estabilización.
- **A las personas responsables funcionales de los procesos afectados.** Debe informarse de que la transición del procedimiento alternativo a la operación con el sistema de IA va a producirse, de las condiciones y el calendario de la transición y de las medidas de supervisión que se aplicarán.
- **Al proveedor del sistema, cuando proceda**. Debe informarse al proveedor de la reactivación del sistema y de las medidas de monitorización reforzada establecidas, para que pueda ajustar su nivel de soporte durante el período de estabilización conforme a los acuerdos contractuales.
- **A las autoridades competentes, cuando proceda**. Cuando las notificaciones realizadas durante el incidente lo requieran, debe comunicarse a las autoridades pertinentes la reactivación del sistema y el estado de las acciones correctoras.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

9.8. Documentación y lecciones aprendidas

La documentación del incidente y el proceso de lecciones aprendidas constituyen la fase final de la respuesta. Su objetivo es doble: por un lado, generar un registro completo y riguroso del incidente que satisfaga las necesidades operativas, legales y regulatorias de la organización; por otro, extraer del incidente el conocimiento necesario para mejorar las capacidades de prevención, detección y respuesta de la organización ante futuros incidentes.

La documentación no es una actividad que deba realizarse únicamente al final del proceso. A lo largo de todas las fases anteriores se ha insistido en la necesidad de documentar las decisiones adoptadas, las acciones ejecutadas, las evidencias preservadas y los resultados obtenidos. La actividad descrita en este apartado se refiere a la consolidación, revisión y formalización de toda esa documentación en un expediente completo del incidente, y a la realización del análisis posterior orientado a la mejora continua.

9.8.1. Expediente del incidente

La organización debe elaborar un expediente completo de cada incidente que afecte a un sistema de IA. Este expediente debe contener, al menos, los siguientes elementos:

Categoría	Contenido
Identificación	Identificador único del incidente. Sistema o sistemas de IA afectados. Fecha y hora de detección. Fuente de detección. Fecha y hora estimadas de inicio del compromiso. Fecha y hora de resolución.
Clasificación	Clasificación inicial y clasificaciones posteriores si hubo reclasificación, con justificación de cada cambio. Tipología del incidente conforme a la sección 6. Severidad asignada. Componentes afectados. Dimensiones de seguridad afectadas.
Cronología	Secuencia detallada de todos los eventos relevantes, desde el inicio estimado del compromiso hasta el cierre del incidente, incluyendo las fechas y horas de cada acción de detección, contención, análisis, erradicación y recuperación.
Análisis	Causa raíz identificada. Vector de ataque. Alcance temporal y funcional del compromiso. Resultados del análisis forense de infraestructura. Resultados del análisis de los componentes de IA. Determinación del período de compromiso y del volumen de inferencias afectadas.
Impacto	Evaluación del impacto sobre los resultados del sistema. Procesos de decisión afectados. Personas afectadas identificadas. Evaluación del impacto sobre servicios. Evaluación de la brecha de datos personales cuando proceda. En entornos de salud, resultados de la revisión clínica. En entornos de Smart Cities, evaluación del impacto sobre el servicio público y la ciudadanía.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

Acciones ejecutadas	Opción de contención adoptada y justificación. Acciones de contención ejecutadas sobre infraestructura, componentes de IA y procesos dependientes. Acciones de erradicación ejecutadas por componente. Resultados de la verificación de la erradicación. Condiciones y fecha de reactivación del sistema. Medidas de monitorización reforzada aplicadas durante el período de estabilización.
Decisiones	Todas las decisiones relevantes adoptadas durante la gestión del incidente, con identificación de la persona que adoptó cada decisión, la fecha, la información disponible en el momento y la justificación. Incluye la decisión de contención, la decisión de reactivación y, cuando proceda, las decisiones sobre revisión de decisiones clínicas o de servicio público.
Notificaciones	Notificaciones realizadas a autoridades competentes, con fechas, contenido y destinatarios. Notificaciones a personas afectadas cuando proceda. Notificación al proveedor. Comunicaciones internas realizadas.
Evidencias	Inventario de todas las evidencias preservadas, con referencia a su ubicación, su cadena de custodia y las condiciones de su conservación.

TABLA 16: CONTENIDO DEL EXPEDIENTE DEL INCIDENTE

El expediente debe elaborarse con el nivel de rigor y detalle suficiente para cumplir las siguientes finalidades:

- Servir como registro operativo que permita a la organización reconstruir con exactitud lo ocurrido y las acciones adoptadas.
- Satisfacer las obligaciones de documentación exigidas por el marco normativo aplicable, incluyendo las obligaciones derivadas del RGPD en relación con las violaciones de seguridad de datos personales, las obligaciones del RIA para los sistemas de IA de alto riesgo, las obligaciones del ENS y las obligaciones derivadas de la Directiva NIS2 cuando resulten aplicables.
- Servir como base probatoria en eventuales procedimientos legales, administrativos o sancionadores.
- Alimentar el proceso de lecciones aprendidas descrito en el apartado siguiente.

El expediente debe conservarse durante un período que permita el cumplimiento de todas las obligaciones normativas aplicables y que tenga en cuenta los plazos de prescripción de las posibles responsabilidades derivadas del incidente. Cuando no exista un plazo específico establecido por la normativa sectorial aplicable, se recomienda un período mínimo de conservación de cinco años desde la fecha de cierre del incidente, sin perjuicio de que la evaluación de las circunstancias concretas del incidente pueda aconsejar un período superior.

9.8.2. Lecciones aprendidas

Una vez cerrado el incidente y completado el expediente, la organización debe realizar un análisis estructurado orientado a identificar las lecciones aprendidas y a traducirlas en mejoras concretas de sus capacidades de prevención, detección y respuesta.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

Este análisis debe realizarse con la participación de todos los perfiles que intervinieron en la gestión del incidente: equipos de ciberseguridad, equipos de IA y ciencia de datos, personas responsables funcionales de los procesos afectados, Delegado de Protección de Datos cuando haya intervenido y, cuando resulte pertinente, representantes del proveedor del sistema.

El análisis debe abordar, al menos, las siguientes áreas:

- **Preparación.** Adecuación del inventario del sistema y de los baselines documentados. Operatividad y eficacia de los procedimientos alternativos. Suficiencia de los acuerdos con proveedores. Disponibilidad de las capacidades técnicas necesarias para el análisis de componentes de IA.
- **Detección**. Eficacia de los mecanismos de monitorización específicos de IA. Tiempo transcurrido entre el inicio del compromiso y su detección. Funcionamiento de los canales de reporte del personal profesional. Necesidad de incorporar nuevos indicadores de compromiso o fuentes de detección.
- **Clasificación y contención.** Adecuación de la clasificación inicial. Idoneidad y rapidez de la opción de contención adoptada. Eficacia de la coordinación entre equipos. Corrección en la preservación de evidencias.
- **Análisis y erradicación.** Grado de confianza en la identificación de la causa raíz. Precisión en la determinación del alcance temporal y funcional. Completitud de la erradicación y adecuación de su verificación.
- **Recuperación.** Adecuación de los criterios de reactivación. Eficacia del período de estabilización. Alcance y profundidad de la revisión de decisiones. Adecuación de la comunicación de la reactivación.
- **Coordinación y comunicación**. Funcionamiento del modelo de coordinación definido en la sección 8. Eficacia de los mecanismos de comunicación interna y externa. Cumplimiento de las obligaciones de notificación en los plazos establecidos.

Los resultados del análisis de lecciones aprendidas deben traducirse en un plan de acciones de mejora concreto, con responsables asignados, plazos de ejecución y mecanismos de seguimiento. Las mejoras pueden afectar a cualquiera de los elementos del proceso de respuesta: actualización del inventario de sistemas, refuerzo de las capacidades de monitorización, modificación de los procedimientos de respuesta, revisión de los acuerdos con proveedores, mejora de la formación de los equipos, actualización de los procedimientos alternativos o refuerzo de los controles de seguridad del sistema.

El plan de acciones de mejora debe integrarse en los procesos de gestión de la seguridad de la organización y su ejecución debe ser objeto de seguimiento hasta su completitud.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

10. Obligaciones de notificación

Cuando un incidente de ciberseguridad afecta a un sistema de IA, pueden activarse simultáneamente obligaciones de notificación derivadas de diferentes instrumentos normativos. Esta sección identifica las principales obligaciones de notificación aplicables, sus condiciones de activación y sus particularidades cuando el incidente afecta a un sistema de IA. No pretende sustituir el análisis jurídico que cada organización debe realizar a la vista de las circunstancias concretas de cada incidente, sino proporcionar un marco de referencia que facilite la identificación temprana de las obligaciones aplicables y la planificación de su cumplimiento.

10.1. Marco general de obligaciones

La siguiente tabla resume las principales obligaciones de notificación que pueden resultar aplicables ante un incidente en un sistema de IA, identificando para cada una el instrumento normativo del que se deriva, el criterio de activación, el destinatario, el plazo y las particularidades más relevantes cuando el incidente afecta a un sistema de IA:

Normativa	Criterio de activación	Destinatario	Plazo	Particularidades en incidentes de IA
RGPD (arts. 33-34)	Violación de seguridad de datos personales con riesgo para los derechos y libertades de las personas físicas	AEPD. Personas afectadas cuando el riesgo sea alto.	72 horas (autoridad). Sin dilación indebida (personas afectadas).	La violación puede producirse a través del modelo (memorización, inversión, fuga en resultados) y no solo mediante acceso a bases de datos. El período de exposición puede ser prolongado. Los datos afectados pueden incluir categorías especiales.
RIA (art. 73)	Incidente grave en sistema de IA de alto riesgo: incidente o defecto que haya provocado o pueda provocar muerte, daños graves a la salud, la seguridad o los derechos fundamentales	Autoridad de vigilancia de mercado competente	Sin dilación indebida, conforme a los plazos que establezca la normativa de desarrollo	La evaluación de gravedad debe considerar el impacto algorítmico y el carácter acumulativo del daño derivado de decisiones erróneas durante períodos prolongados.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

ENS (art. 36, CCN-STIC 817)	Incidente de seguridad en sistemas del ámbito del ENS	CCN-CERT	Conforme a los procedimientos del CCN-CERT según el nivel de peligrosidad	La clasificación debe reflejar las especificidades del compromiso de IA. La información reportada debe incluir los componentes afectados y el impacto sobre los resultados del sistema.
Directiva NIS2 (arts. 23-24)	Incidente significativo en entidad esencial o importante	CSIRT de referencia y autoridad competente	Alerta temprana: 24 horas. Notificación: 72 horas. Informe final: un mes.	La afectación al servicio puede derivarse del impacto algorítmico sin interrupción técnica del sistema.
MDR/IVDR	Incidente grave en producto sanitario o producto sanitario para diagnóstico in vitro	AEMPS	Entre 2 y 15 días según gravedad	Aplicable cuando el sistema de IA esté calificado como producto sanitario. Debe considerarse el impacto clínico de los resultados comprometidos.

TABLA 17: OBLIGACIONES DE NOTIFICACIÓN APLICABLES A INCIDENTES EN SISTEMAS DE IA

10.2. Consideraciones específicas para incidentes en sistemas de IA

La aplicación de las obligaciones de notificación a los incidentes que afectan a sistemas de IA presenta un conjunto de particularidades que deben tenerse en cuenta para un cumplimiento adecuado:

- **Determinación del momento de conocimiento.** Los plazos se computan desde el conocimiento del incidente, cuya determinación puede no ser evidente en incidentes con impacto algorítmico. La organización no debe utilizar esta incertidumbre para demorar la notificación: cuando existan indicios razonables de violación de seguridad o incidente grave, el plazo debe considerarse activado, sin perjuicio de que la notificación inicial se complete con información adicional a medida que avance la investigación.
- **Concurrencia de obligaciones.** Un mismo incidente puede activar simultáneamente obligaciones conforme al RGPD, al RIA, al ENS, a la Directiva NIS2 y al MDR. La organización debe identificar desde las fases iniciales todas las obligaciones potencialmente aplicables y planificar su cumplimiento de forma coordinada.
- **Evaluación del impacto sobre datos personales.** La evaluación conforme al RGPD debe considerar las vías específicas de afectación a datos personales en sistemas de IA, que no se limitan al acceso a bases de datos, sino que incluyen la extracción mediante técnicas de inversión del modelo, la memorización y revelación de datos de entrenamiento por modelos generativos, la inferencia de información personal a partir de resultados y la fuga de datos a través de las respuestas del sistema.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD
Junta de Andalucía
Agencia Digital de Andalucía

- **Contenido de las notificaciones**. Además de la información requerida por cada instrumento normativo, las notificaciones deben incluir la identificación del sistema de IA afectado y su clasificación regulatoria, la tipología del incidente, los componentes comprometidos, el período estimado de compromiso y el impacto sobre resultados y decisiones. Las notificaciones a personas afectadas deben formularse en lenguaje claro, centrado en las consecuencias prácticas y en las medidas de protección disponibles.
- **Notificación por fases.** La investigación de incidentes en sistemas de IA puede requerir plazos prolongados, especialmente para la determinación del alcance temporal y la evaluación del impacto sobre decisiones. Las normativas permiten, con carácter general, la notificación por fases: notificación inicial dentro del plazo establecido con la información disponible, completada progresivamente a medida que avance el análisis. Esta posibilidad no exime del cumplimiento del plazo inicial.

10.3. Coordinación de las notificaciones

La organización debe designar una persona responsable de la coordinación de las notificaciones derivadas de cada incidente, que se encargue de identificar todas las obligaciones aplicables, verificar el cumplimiento de los plazos, asegurar la coherencia del contenido de las diferentes notificaciones y documentar todas las comunicaciones realizadas. Esta función de coordinación debe ejercerse en colaboración con el Delegado de Protección de Datos, con el servicio jurídico de la organización y con los equipos técnicos que proporcionan la información sobre el incidente.

Cuando el incidente afecte a un sistema proporcionado por un proveedor externo, debe coordinarse con el proveedor para determinar las obligaciones de notificación que correspondan a cada parte conforme al marco normativo y a los acuerdos contractuales, evitando tanto duplicidades como omisiones en las notificaciones.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

11. Referencias

- Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (AI Act). https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32024R1689

- Reglamento (UE) 2016/679 (RGPD), de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos. https://eur-lex.europa.eu/eli/reg/2016/679/oj

- Ley Orgánica 3/2018, de 5 de diciembre (LOPDGDD), de Protección de Datos Personales y garantía de los derechos digitales. https://www.boe.es/buscar/act.php?id=BOE-A-2018-16673

- Directiva (UE) 2022/2555 (NIS2), relativa a medidas para un elevado nivel común de ciberseguridad en la Unión. https://eur-lex.europa.eu/eli/dir/2022/2555/oj

- Real Decreto 311/2022, de 3 de mayo, por el que se regula el Esquema Nacional de Seguridad (ENS). https://www.boe.es/eli/es/rd/2022/05/03/311

- Reglamento (UE) 2024/2847 (Cyber Resilience Act – CRA), relativo a los requisitos horizontales de ciberseguridad para productos con elementos digitales. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32024R2847

- ISO/IEC 42001:2023, Artificial Intelligence — Management system. https://www.iso.org/standard/81230.html

- ISO/IEC 23894:2023, Artificial intelligence — Risk management. https://www.iso.org/standard/77304.html

- ISO/IEC 27001:2022, Information security management systems — Requirements. https://www.iso.org/standard/82875.html

- NIST AI Risk Management Framework 1.0 (2023). https://www.nist.gov/itl/ai-risk-management-framework

12. Otras referencias de interés

El presente apartado recoge un resumen de los principales temas abordados en la guía, así como la relación de productos y servicios mencionados en ella. Su finalidad es ofrecer una visión global de los ámbitos tratados y facilitar la identificación de posibles referencias o elementos relevantes para futuras consultas o ampliaciones documentales.

12.1. Lista de materias tratadas en la guía

- Particularidades de los incidentes de ciberseguridad en sistemas de IA frente a incidentes convencionales
- Componentes vulnerables de un sistema de IA y su impacto en la gestión de incidentes
- Tipología de incidentes de ciberseguridad específicos de sistemas de IA
- Principios de respuesta a incidentes en sistemas de IA
- Modelo organizativo de respuesta: roles, responsabilidades y mecanismos de coordinación
- Preparación: inventario de sistemas, capacidades de detección y acuerdos con proveedores
- Detección e identificación: fuentes, indicadores de compromiso específicos de IA y distinción entre degradación natural y compromiso malicioso
- Clasificación, priorización y evaluación de severidad de incidentes en sistemas de IA
- Contención: opciones de contención, preservación de evidencias y gestión de procesos dependientes
- Análisis e investigación: análisis forense de componentes de IA, determinación del alcance temporal y evaluación del impacto sobre personas y servicios
- Erradicación: restauración de modelos, datos, pipelines y verificación de la erradicación
- Recuperación: criterios de reactivación, reintegración progresiva y revisión de decisiones adoptadas durante el período de compromiso
- Documentación, expediente del incidente y lecciones aprendidas
- Obligaciones de notificación: marco normativo aplicable y coordinación de notificaciones

12.2. Lista de productos mencionados en la guía

No se mencionan productos específicos en la guía.

12.3. Lista de servicios mencionados en la guía

No se mencionan productos específicos en la guía.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales
Plan de Recuperación, Transformación y Resiliencia
incibe_ Instituto Nacional de Ciberseguridad
Junta de Andalucía
Agencia Digital de Andalucía

RETECH Redes Territoriales de Especialización Tecnológica

RED ARGOS

Guía de respuesta a incidentes de ciberseguridad en sistemas IA en entornos de Salud y Smart Cities

Febrero de 2026

Versión 1.0

Junta de Andalucía

Financiado por la Unión Europea NextGenerationEU

Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Telecomunicaciones e Infraestructuras Digitales

Plan de Recuperación, Transformación y Resiliencia

incibe_ Instituto Nacional de Ciberseguridad

Junta de Andalucía

Agencia Digital de Andalucía