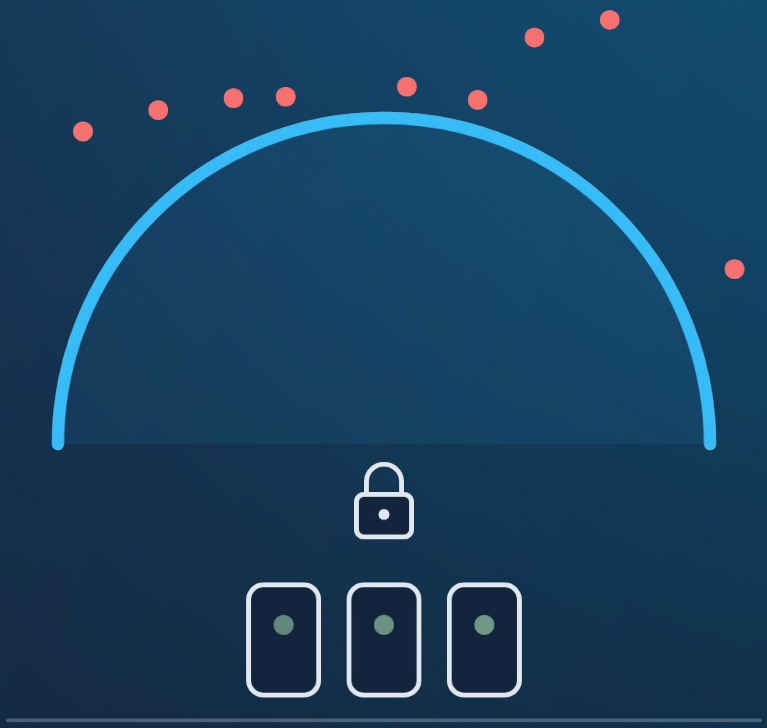
CLOUD LEGION



EBOOK · GUÍA PRÁCTICA

Seguridad de la IA con ISO/IEC 27090

Las nueve amenazas que pueden comprometer sus sistemas de IA y cómo detectarlas y mitigarlas en todo el ciclo de vida.

cloudlegion.com.ar

CONTENIDO

Lo que encontrará en esta guía



La ISO/IEC 27090 es la primera norma internacional dedicada a las amenazas propias de los sistemas de IA. Esta guía la explica en lenguaje claro y la baja a un plan concreto.

01 **Por qué ahora** — La IA se puede atacar sin tocar la infraestructura — 03

02 **Qué es la ISO/IEC 27090** — Alcance, estructura y la base de la cláusula 5 — 04

03 **Las nueve amenazas** — Los tipos de compromiso de la cláusula 6, con ejemplos — 08

04 **Ciclo de vida, mitigaciones y agentes** — Por qué la producción también entrena — 22

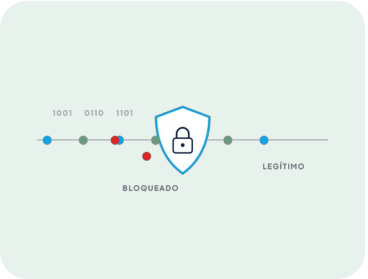
05 **Cómo encaja con otras normas** — Mapa de controles de ISO 27002 e ISO 42001, y la Ley de IA — 31

06 **Llevarla a la práctica** — Preguntas clave, roles, un caso y una hoja de ruta de 90 días — 34

07 **Cómo puede ayudar Cloud Legion** — Servicios para proteger sus sistemas de IA — 41

01 · POR QUÉ AHORA

La IA se puede atacar sin tocar la infraestructura



Un modelo de IA puede ser comprometido aunque el firewall, los servidores y las contraseñas estén en orden. Alguien puede contaminar sus datos de entrenamiento, copiarlo con consultas que parecen legítimas o hacerle obedecer instrucciones escondidas en un documento.

Estos ataques apuntan a lo que la IA tiene de distinto: sus datos, sus modelos y lo que produce. Los controles clásicos siguen siendo necesarios, pero no alcanzan. Por eso ISO e IEC publicaron una norma específica.

14/10/2026

publicación de la primera edición

56

páginas, del comité ISO/IEC JTC 1/SC 27

9

tipos de compromiso propios de la IA

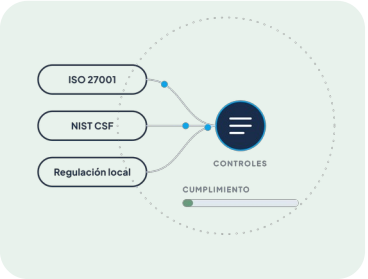
2

obligaciones formales: el resto es guía

LA IDEA CLAVE

La seguridad de la IA no termina en el despliegue. Los datos que el sistema recibe en producción pueden terminar reentrenando el modelo: un atacante puede envenenarlo sin tocar nunca el entorno de desarrollo.

Una guía para proteger la IA, no un certificado



ALCANCE

Ayuda a entender las consecuencias de las amenazas específicas de los sistemas de IA en todo su ciclo de vida, y describe cómo detectarlas y mitigarlas. Aplica a organizaciones de cualquier tipo y tamaño que desarrollen o usen IA: no solo chatbots, también modelos de scoring, visión por computadora, recomendadores o detección de fraude.

Qué es

- Una guía técnica sobre amenazas y mitigaciones propias de la IA
- Un complemento de ISO/IEC 27001, 27002 y 42001
- Aplicable a quien desarrolla y a quien usa IA de terceros

Qué no es

- Un sistema de gestión ni una norma certificable
- Una prueba de cumplimiento de la Ley de IA europea u otra ley
- Un reemplazo de los controles de seguridad clásicos

FICHA

Nombre	Ciberseguridad — Inteligencia artificial — Tratamiento de amenazas y compromisos de seguridad de los sistemas de IA
Edición	Primera, publicada el 14 de octubre de 2026
Comité	ISO/IEC JTC 1/SC 27 (seguridad de la información, ciberseguridad y privacidad)
Referencias normativas	ISO/IEC 22989 (conceptos de IA) e ISO/IEC 23053 (marco de sistemas de IA con aprendizaje automático)
Anexos informativos	A: ataques por etapa y por activo · B: versiones propias de la IA de ataques clásicos · C: aprendizaje federado

Cómo está organizada la norma



5 — **Aplicación de la seguridad de la información**

Gobernanza, zero trust, ingeniería de software, cadena de suministro, modelado de amenazas y red teaming.

6 — **Tipos de compromiso (6.2 a 6.10)**

Nueve amenazas. Para cada una: descripción, fallas que provoca, detección y mitigación.

6.11 — **Consideraciones para sistemas de IA**

Minimizar datos y detalles técnicos, supervisión, mínimo privilegio para agentes y procedencia.

7 — **Mitigaciones y sus interacciones**

Conflictos entre controles, continuidad en el ciclo, deterioro, registro, entornos y filtrado.

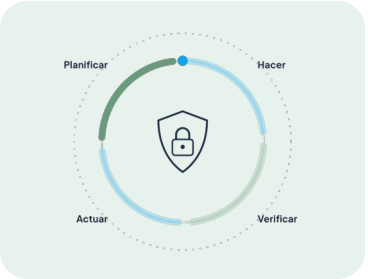
A–C — **Anexos informativos**

Mapa de ataques por etapa y activo, ataques clásicos en versión IA y aprendizaje federado.

LAS DOS OBLIGACIONES DE LA CLÁUSULA 5.1

Solo dos: implementar controles de seguridad, como los de ISO/IEC 27002, según una evaluación de riesgos; y tener en cuenta los compromisos propios de la IA en todo el ciclo de vida.

Qué hay que proteger en un sistema de IA



Antes de hablar de amenazas, hay que saber qué proteger. El Anexo A identifica los activos de un sistema de aprendizaje automático y la cláusula 6 ordena cada amenaza según el activo que compromete primero y la propiedad que afecta: confidencialidad, integridad o disponibilidad.

Datos de entrenamiento y prueba	Las muestras y sus etiquetas. Base del aprendizaje, la evaluación y las pruebas.
Algoritmo o modelo	La transformación que convierte entradas en predicciones, con su arquitectura.
Parámetros entrenados	Los pesos aprendidos, presentes en el entrenamiento y en la inferencia.
Salida del modelo	La predicción o decisión que el sistema entrega.
Herramientas de preparación	Lo que procesa los datos antes de enviarlos al modelo.

IMPACTOS QUE MAPEA LA NORMA

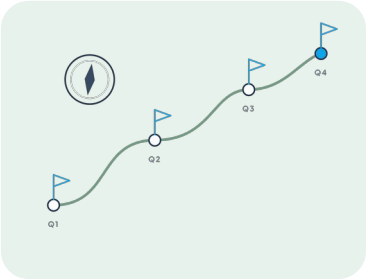
T	B	P	A	L
Confidencialidad de los datos de entrenamiento	Comportamiento del modelo	Propiedad intelectual	Disponibilidad	Confidencialidad de las entradas

POR QUÉ IMPORTA

Los marcos de control tradicionales siguen valiendo, pero hay que aplicarlos por separado a cada etapa del pipeline de ML: datos, entrenamiento, modelo, inferencia y salidas.

02 · CLÁUSULA 5

La base: seguridad de la información aplicada a la IA



Antes de hablar de ataques exóticos, la norma pide lo básico bien hecho: un sistema de IA es software integrado en otros sistemas, y todo lo que protege al software también lo protege a él.

5.2 · Gobernanza de IA

Inventario de los sistemas de IA y su uso, y análisis y gestión de sus riesgos.

5.3 · Zero trust

Asumir la brecha, verificar cada pedido y dar el mínimo privilegio, también con el modelo.

5.4 · Ingeniería de software

Equipos que unen ciencia de datos y desarrollo; trazabilidad de datos y modelos.

5.5 · Cadena de suministro

Troyanos en modelos, dependencias maliciosas, procedencia falseada. Firmar y verificar.

5.6 · Modelado de amenazas

Usar los escenarios de la cláusula 6 para encontrar superficies y vectores de ataque.

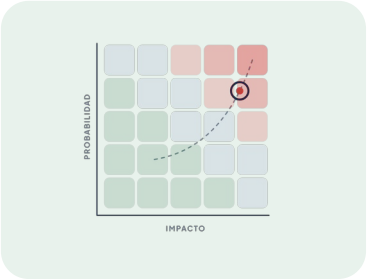
5.7 · Red teaming

Emular a adversarios reales para encontrar las vulnerabilidades antes que ellos.

UN CONSEJO DE LA NORMA

Asegurar primero la base de software, hardware y red suele ser lo más eficiente: contra el robo del modelo, rinde más proteger el entorno que defenderse de la exfiltración por consultas.

Los tipos de compromiso de la cláusula 6



Ejemplos nuestros; mitigaciones resumidas de la norma.

Atacan los datos y lo que el modelo aprende

6.2 Envenenamiento de datos

QUÉ ES Se alteran los datos de entrenamiento o, si el modelo aprende de forma continua, los de entrada.

EJEMPLO Transacciones fraudulentas cargadas como legítimas para que el detector de fraude aprenda a ignorarlas.

MITIGACIÓN Asumir que los datasets pueden estar comprometidos: sanitizarlos, detectar anomalías y deriva, y conocer su procedencia.

6.3 Evasión

QUÉ ES Entradas alteradas, a menudo imperceptibles para una persona, que hacen fallar al modelo.

EJEMPLO Un documento de identidad retocado que pasa un control automático de onboarding.

MITIGACIÓN Entrenamiento adversarial y detección de entradas adversariales o fuera de distribución.

6.4 Inferencia de pertenencia

QUÉ ES Deducir si un registro concreto formó parte de los datos de entrenamiento.

EJEMPLO Confirmar que un paciente estaba en el dataset de un modelo clínico.

MITIGACIÓN Privacidad diferencial, desidentificación, auditoría de los datos y filtrado de entradas y salidas.

Atacan el modelo y lo que sabe

6.5 Exfiltración del modelo

QUÉ ES Recrear un modelo con la misma funcionalidad a partir de consultas válidas.

EJEMPLO Miles de llamadas a una API de scoring para entrenar un clon de su lógica.

MITIGACIÓN No exponer puntajes de probabilidad, limitar consultas, marcas de agua y detectar patrones de consulta inusuales.

6.6 Inversión del modelo

QUÉ ES Usar las salidas para reconstruir información sensible de los datos de entrenamiento.

EJEMPLO Recuperar rasgos de los rostros usados para entrenar un sistema de reconocimiento.

MITIGACIÓN Limitar la cantidad y frecuencia de consultas, privacidad diferencial y datos sintéticos o desidentificados.

6.7 Envenenamiento directo del modelo

QUÉ ES Alterar el modelo o su código y algoritmo de entrenamiento, sin tocar los datos.

EJEMPLO Un modelo descargado de un repositorio público que trae una puerta trasera.

MITIGACIÓN Proteger los entornos de desarrollo y ejecución, gestionar la cadena de suministro, cifrar y microsegmentar.

POR QUÉ IMPORTA

Robar un modelo o hacerle revelar datos no requiere entrar a la red: alcanza con usarlo como lo usaría un cliente.

Atacan lo que el sistema produce

6.8 Salida sensible

QUÉ ES La salida revela datos de entrenamiento o de otros usuarios. Puede ocurrir sin atacante.

EJEMPLO Un asistente que responde con datos de otro cliente que el modelo memorizó.

MITIGACIÓN Detección automática de datos sensibles en las salidas y varias capas de revisión antes de entregarlas.

6.9 Inyección de prompts

QUÉ ES Instrucciones en la entrada provocan un comportamiento no previsto. Incluye el jailbreaking.

EJEMPLO Un correo con texto oculto que ordena al asistente reenviar archivos a un tercero.

MITIGACIÓN Separar la entrada no confiable, sanitizar el contenido recuperado (RAG), mínimo privilegio y límites de uso.

6.10 Salida que contiene inyecciones

QUÉ ES La respuesta del modelo lleva un ataque, como cross-site scripting, para el sistema que la consume.

EJEMPLO Código generado que incluye una consulta SQL maliciosa o un script en una página web.

MITIGACIÓN Validar y codificar las salidas como cualquier dato no confiable antes de ejecutarlas o mostrarlas.

EL DATO

La inyección de prompts es la amenaza número uno del OWASP Top 10 para aplicaciones con LLM desde su primera edición. En la versión 2025, la divulgación de información sensible subió al segundo lugar.

Cuando el modelo aprende lo que el atacante quiere



CÓMO FUNCIONA

El atacante inyecta datos corruptos en el set de entrenamiento, o reordena los existentes mientras el modelo se entrena. Si la validación y la prueba salen del mismo dataset, también quedan contaminadas. Y como muchos sistemas se apoyan en modelos fundacionales entrenados con datos públicos, el ataque ni siquiera necesita entrar a la organización.

Puede ser dirigido, con un activador que provoca errores en casos puntuales, o indiscriminado, para destruir la integridad del dataset.

QUÉ PUEDE FALLAR

- Decisiones equivocadas a partir de predicciones falsas
- Integridad comprometida y, en cascada, confidencialidad y disponibilidad
- Tiempo fuera de servicio para investigar y remediar

CÓMO SE DETECTA

- Es un campo de investigación abierto: cuesta distinguir un dato malicioso de uno atípico
- Controles de integridad sobre los datasets en todo su ciclo: altas, bajas y cambios
- Monitoreo continuo de datasets, modelos y sistemas

CÓMO SE MITIGA

- Asumir que los datasets pueden estar comprometidos y sanitizarlos antes de usarlos
- Detectar anomalías y medir sesgos o deriva en la distribución de los datos
- Conocer la procedencia, la frecuencia y las fuentes de cada actualización

PREGUNTA PARA SU EQUIPO

¿Sabemos quién cambió cada dataset y cuándo?

03 · 6.3 EVASIÓN

Entradas que engañan al modelo, no a las personas



CÓMO FUNCIONA

El atacante introduce pequeñas perturbaciones en una entrada que el modelo clasificaba bien. En visión, una imagen que una persona ve idéntica; en ciberseguridad, un malware modificado que conserva su lógica pero que el detector clasifica como seguro.

La norma ordena los errores en tres niveles de impacto creciente: fallas en entornos conocidos, situaciones poco frecuentes que el entrenamiento no cubrió y ataques deliberados, físicos o digitales.

QUÉ PUEDE FALLAR

- Falsos positivos o falsos negativos en clasificadores
- Comportamientos con consecuencias para la seguridad física, como en vehículos autónomos
- Pérdida de disponibilidad del sistema que integra el modelo

CÓMO SE DETECTA

- Difícil de advertir a simple vista
- Redes complementarias que evalúan si cada entrada es adversarial
- Análisis de la distribución de entradas y salidas en el tiempo

CÓMO SE MITIGA

- Entrenamiento adversarial, sin olvidar sus efectos sobre otras defensas (cláusula 7)
- Detección de entradas adversariales y fuera de distribución
- Evaluación de robustez con ISO/IEC TR 24029-1 e ISO/IEC 24029-2

PREGUNTA PARA SU EQUIPO

¿Probamos el modelo con entradas diseñadas a propósito para engañarlo?

Saber que alguien estaba en los datos ya es una fuga



CÓMO FUNCIONA

Los modelos suelen comportarse distinto con los datos que vieron durante el entrenamiento. El atacante aprovecha esa diferencia para deducir si un registro concreto formó parte del dataset. El ataque funciona mejor en modelos que generalizan mal y memorizan casos puntuales.

El ejemplo de la norma es claro: en un sistema de salud, confirmar que una persona estaba en los datos de entrenamiento puede revelar que tiene una determinada patología.

QUÉ PUEDE FALLAR

- Compromiso de la confidencialidad y la privacidad de los datos
- Con datos personales, un incidente de privacidad
- Obligaciones de notificación según la jurisdicción

CÓMO SE DETECTA

- Difícil, porque las consultas parecen normales
- Analizar frecuencia, secuencia, atributos, horarios y diversidad de las consultas
- Usar lo detectado para reforzar el entrenamiento, por ejemplo quitando atributos muy correlacionados

CÓMO SE MITIGA

- Privacidad diferencial, sabiendo que rara vez logra un buen equilibrio entre utilidad y privacidad
- Desidentificación de los datos de entrenamiento
- Auditoría de los datos y filtrado de entradas y salidas, coordinados entre sí

PREGUNTA PARA SU EQUIPO

¿Entrenamos con datos personales innecesarios?

Copiar un modelo sin tocar un solo archivo



CÓMO FUNCIONA

Con muchas consultas válidas, el atacante recrea un modelo con la misma funcionalidad que el original. Es un ataque de oráculo: alcanza con preguntar. No hay que confundirlo con el robo directo del modelo (Anexo B), que copia pesos o estructura entrando al sistema.

El daño depende de cuán novedoso sea el modelo y de cuánto costó entrenarlo.

QUÉ PUEDE FALLAR

- Pérdida de propiedad intelectual
- El atacante usa el modelo para sus propios fines
- Abre la puerta a la inversión del modelo y a la fuga de datos de entrenamiento

CÓMO SE DETECTA

- Analítica de comportamiento: muchas consultas desde el mismo origen
- Umbrales según cuánto conocimiento pudo obtener el atacante
- Consultas demasiado uniformes, demasiado aleatorias o centradas en atributos puntuales

CÓMO SE MITIGA

- No exponer puntajes de probabilidad ni detalles innecesarios
- Limitar la frecuencia de consultas y bloquear orígenes sospechosos
- Marcas de agua para demostrar la autoría del modelo

PREGUNTA PARA SU EQUIPO

¿Nos daríamos cuenta si alguien hace miles de consultas sistemáticas a nuestra API?

Reconstruir datos sensibles a partir de las respuestas



CÓMO FUNCIONA

El atacante busca las entradas para las que el modelo responde con confianza muy alta y así recupera rasgos privados de los datos de entrenamiento. Está demostrado con imágenes y texto: se llegaron a reconstruir rostros de un sistema de reconocimiento facial, y los modelos de lenguaje pueden filtrar datos memorizados como códigos postales o números de identificación.

QUÉ PUEDE FALLAR

- Compromiso permanente de la confidencialidad de los datos de entrenamiento
- Exposición de información de personas concretas
- Consecuencias de privacidad que no se pueden revertir

CÓMO SE DETECTA

- Las mismas técnicas que para la exfiltración
- Muchas consultas desde el mismo origen
- Patrones que recorren todo el espacio de predicciones

CÓMO SE MITIGA

- Limitar la cantidad y la frecuencia de consultas
- Privacidad diferencial
- Desidentificación o datos sintéticos durante el entrenamiento

PREGUNTA PARA SU EQUIPO

¿Nuestra API devuelve más información de la que el cliente realmente necesita?

El problema no está en los datos, sino en el modelo



CÓMO FUNCIONA

Se manipula el comportamiento sin tocar los datos: alterando el código del pipeline de entrenamiento, entregando un modelo o una arquitectura envenenados, o modificando el modelo almacenado durante el desarrollo o la operación.

Si un proveedor entrena el modelo de forma manipulada y la organización lo reutiliza o lo ajusta, el comportamiento indeseado sobrevive: es lo que la norma llama un ataque de transferencia de aprendizaje.

QUÉ PUEDE FALLAR

- Las mismas que el envenenamiento de datos: decisiones erróneas e integridad comprometida
- El defecto puede sobrevivir al ajuste fino
- Difícil de atribuir cuando viene de terceros

CÓMO SE DETECTA

- Como parte de la detección general de ataques, si ocurre en desarrollo u operación
- Inteligencia de amenazas sobre fuentes y modelos maliciosos conocidos

CÓMO SE MITIGA

- Proteger el entorno de desarrollo y el de operación, por ejemplo con entornos de ejecución confiables
- Gestión de la cadena de suministro y las mismas medidas que contra el envenenamiento de datos
- Microsegmentación, cifrado en reposo, en tránsito y en uso, y revisión antes de desplegar

PREGUNTA PARA SU EQUIPO

¿Verificamos cada modelo antes de desplegarlo?

Cuando el modelo dice lo que no debía



CÓMO FUNCIONA

La salida contiene datos sensibles del entrenamiento, por ejemplo datos personales que un modelo de lenguaje memorizó. Un caso particular es el texto o las imágenes protegidos por derechos de autor.

La diferencia con otras amenazas es que puede ocurrir sin atacante: aparece por uso normal o porque alguien lo provoca a propósito.

QUÉ PUEDE FALLAR

- Se compromete la confidencialidad de los datos de entrenamiento
- Riesgo legal por datos personales o propiedad intelectual de terceros

CÓMO SE DETECTA

- Filtrar las salidas buscando datos reconocibles por su estructura o contenido, como un número de teléfono

CÓMO SE MITIGA

- Censura automática de los datos sensibles detectados
- Varias capas de revisión de las salidas
- Verificar de forma explícita que no se filtre información del modelo

PREGUNTA PARA SU EQUIPO

¿Revisamos que las respuestas no incluyan datos personales antes de mostrarlas?

Instrucciones escondidas donde nadie mira



CÓMO FUNCIONA

El sistema no distingue entre los datos que recibe y sus propias instrucciones. En la inyección directa, el usuario escribe algo que anula las políticas del sistema; el jailbreaking es una variante. En la indirecta, las instrucciones llegan escondidas en una fuente externa, como una página web que la aplicación agrega al contexto del modelo sin que el usuario lo sepa.

QUÉ PUEDE FALLAR

- Integridad y confidencialidad comprometidas
- Exposición de las instrucciones internas del sistema
- Contenido dañino o acciones no previstas cuando la salida alimenta otro sistema

CÓMO SE DETECTA

- Clasificadores de contenido dañino en entradas y salidas, incluso usando otro modelo de lenguaje
- Registro y monitoreo con heurísticas de comportamiento adversarial
- Listas de frases bloqueadas, con efectividad limitada

CÓMO SE MITIGA

- Separar la entrada no confiable con delimitadores y aclarar en las instrucciones que es solo información
- Sanitizar el contenido recuperado en sistemas RAG
- Limitar accesos y cantidad de consultas, y entrenar con ejemplos adversariales

PREGUNTA PARA SU EQUIPO

¿Qué podría hacer nuestro asistente si un documento le diera una orden?

La respuesta del modelo como vector de ataque



CÓMO FUNCIONA

La salida de texto de un modelo puede traer ataques de inyección clásicos, como cross-site scripting. Si un navegador u otro intérprete procesa esa salida como comandos, el código malicioso se ejecuta.

La regla de la norma es simple: si los datos de entrenamiento no son confiables, la salida tampoco lo es.

QUÉ PUEDE FALLAR

- Ejecución de código malicioso
- Manipulación de la apariencia o del comportamiento de un sitio web

CÓMO SE DETECTA

- Las mismas validaciones y codificaciones estándar de cualquier aplicación

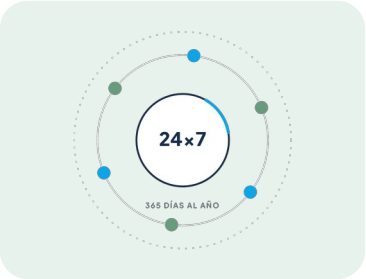
CÓMO SE MITIGA

- Tratar la salida como dato no confiable
- Validar y codificar antes de mostrar o ejecutar
- Varias capas de revisión de las salidas

PREGUNTA PARA SU EQUIPO

¿El código que genera la IA pasa por las mismas revisiones que el escrito por personas?

Ataques clásicos con versión IA



Un sistema de IA sufre los mismos ataques que cualquier sistema de información. El Anexo B muestra los que tienen consideraciones propias de la IA. Primera parte: disponibilidad, cadena de suministro y puertas traseras.

Denegación de servicio

Además de saturar recursos, hay ataques propios de la IA: los ejemplos «esponja» pueden tardar y consumir hasta 30 veces más que una consulta típica.

CÓMO SE ENFRENTA Medir tiempo y energía por consulta desde el diseño; cortar consultas que excedan un presupuesto o dimensionar para el peor caso.

Cadena de suministro

Más allá de los datos, hay componentes críticos: la norma menciona ataques a los generadores de números aleatorios.

CÓMO SE ENFRENTA Contratos con requisitos de seguridad para quienes proveen datos y modelos, capacitación de proveedores, model cards y red teaming.

Puertas traseras

Un tipo de envenenamiento que puede viajar en modelos preentrenados descargados de internet o en servicios de ML.

CÓMO SE ENFRENTA Escanear los archivos de pesos en busca de código malicioso, usar formatos seguros como safetensors y monitorear el comportamiento.

Robos y fugas que no pasan por el modelo



Segunda parte del Anexo B: ataques que no necesitan consultar el modelo, sino llegar a donde están el modelo, los datos de entrenamiento o las entradas de los usuarios.

Robo directo del modelo

Copiar los parámetros entrando al sistema: ejecutables, memoria, almacenamiento o canales laterales.

CÓMO SE ENFRENTA Proteger los entornos de desarrollo y operación; entornos de ejecución confiables.

Fuga de datos de entrenamiento

Los datos reales viven en desarrollo, y al ajustar un modelo en la nube pasan a un tercero.

CÓMO SE ENFRENTA Medidas estándar de protección de datos y detección de pérdida de información.

Fuga de entradas

Lo que los usuarios escriben puede contener datos personales o secretos. La IA generativa tiene más riesgo: entradas abiertas y difíciles de controlar.

CÓMO SE ENFRENTA Cifrado en tránsito, minimizar lo que se envía al modelo y controlar lo que se recupera por RAG.

04 · CICLO DE VIDA

La producción también entrena



Muchos sistemas se reentrenan con lo que reciben en producción. Si esos datos no se controlan, el envenenamiento deja de ser un problema solo del desarrollo.

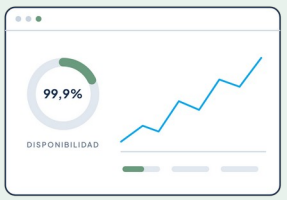
ETAPA	EXPOSICIÓN PRINCIPAL	CONTROLES SUGERIDOS
01 **Recolección de datos**	Datos contaminados o de origen dudoso	Procedencia documentada, validación y control de acceso a los datasets
02 **Desarrollo y entrenamiento**	Envenenamiento, fuga de datos sensibles	Entornos aislados, revisión de datos, pruebas adversariales
03 **Cadena de suministro**	Modelos, datasets y librerías de terceros alterados	Inventario de modelos, verificación de integridad, evaluación de proveedores
04 **Despliegue**	Inyección de prompts, permisos excesivos	Mínimo privilegio, filtrado de entradas, salidas tratadas como no confiables
05 **Operación y monitoreo**	Evasión, exfiltración, salidas sensibles	Registro de consultas, límites de uso, detección de patrones anómalos
06 **Reentrenamiento**	Envenenamiento desde producción	Filtrado y revisión de datos antes de reentrenar, pruebas de regresión

REGLA PRÁCTICA

Los datos nuevos pasan por la misma revisión que los originales, y el modelo reentrenado, por las mismas pruebas que antes del primer despliegue.

04 · CLÁUSULA 7

Los controles interactúan y se desgastan



La cláusula 7 trata algo que suele quedar afuera de las listas de controles: las mitigaciones no son independientes ni eternas. Hay que diseñarlas como un sistema y revisarlas.

01

Las mitigaciones pueden chocar (7.2)

La norma cita pruebas en las que entrenar contra un ataque adversarial volvió al modelo más vulnerable a otro. Y la privacidad diferencial rara vez equilibra bien: si protege mucho, el modelo pierde utilidad.

02

Coherentes en todo el ciclo (7.3)

Contra la inferencia de pertenencia, la auditoría de datos, el entrenamiento con privacidad diferencial y el filtrado en operación deben funcionar juntos.

03

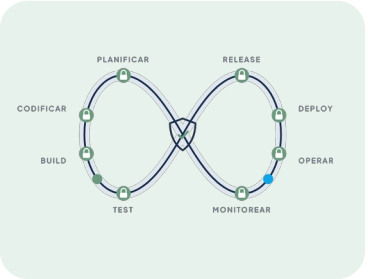
Se degradan con el tiempo (7.4)

Como el rendimiento del modelo, la eficacia de las defensas cae. Hay que monitorearla y validarla otra vez antes de cada redespliegue.

LA DETECCIÓN ES LA OTRA MITAD

Registrar cada uso del modelo con entrada, fecha y hora, usuario, salida y versión (7.5). Buscar series de pequeñas variaciones o consultas sospechosamente sistemáticas (7.7). Y limitar la frecuencia por usuario: no para evitar la sobrecarga, sino para frenar la experimentación del atacante (7.8).

En IA, lo más sensible está en desarrollo



En un sistema tradicional los datos sensibles viven en producción. En IA suele ser al revés: el entorno de desarrollo concentra los datos de entrenamiento, los datos de prueba, el código, las configuraciones, los parámetros del modelo y la documentación técnica. Una fuga ahí expone propiedad intelectual y datos personales, y le facilita el trabajo al atacante.

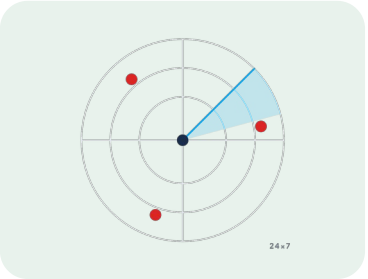
DOS EXCEPCIONES QUE LLEVAN EL PROBLEMA A PRODUCCIÓN

- Modelos que se reentrenan con datos de operación: parte del set de entrenamiento vive en producción.
- Sistemas RAG: el repositorio que se agrega al contexto cumple el papel de datos de entrenamiento y necesita sus mismos controles.

CONTROLES QUE PROPONE, CON SU REFERENCIA EN ISO/IEC 27002

Cifrado de datos en reposo	5.33
Control de acceso con mínimo privilegio	5.15, 5.16, 5.18, 5.3 y 8.3
Seguridad operativa de las estaciones de trabajo	5.23 y 5.37
Monitoreo de manipulaciones sospechosas	8.16
Separación de entornos	8.31
Cómputo confidencial para datos y parámetros en uso	—
Eliminación segura de datos y modelos al final de su vida	—

Filtrar entradas, frenar al atacante y achicar el modelo



01

Entradas fuera de distribución (7.7.1)

Detectar entradas muy distintas de los datos de entrenamiento: pueden indicar un ataque o una deriva. No toda entrada atípica es maliciosa ni todo ataque es atípico, pero registrarlas y, si corresponde, descartarlas ayuda.

02

Entradas adversariales (7.7.2)

Analizar las series de entradas por usuario: pequeñas variaciones con confianza creciente sugieren evasión o inversión; consultas muy sistemáticas sugieren exfiltración. También existen técnicas que distorsionan la entrada para neutralizar la perturbación.

03

Limitar la frecuencia de uso (7.8)

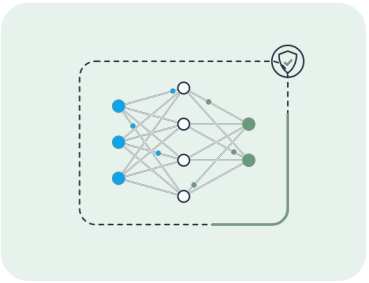
Poner un límite de consultas por usuario. El objetivo no es evitar la sobrecarga, sino demorar la experimentación que estos ataques necesitan.

04

Optimizar el tamaño del modelo (7.9)

Un modelo dimensionado para no memorizar casos individuales dificulta que sobrevivan muestras envenenadas, la inferencia de pertenencia y el detalle de lo que se puede reconstruir por inversión.

Mínimo privilegio y permisos que vencen



Un agente no solo genera texto: consulta bases de datos, envía correos o llama a APIs. Ahí una inyección de prompts deja de ser un problema de contenido y pasa a ser un problema de permisos.

Permisos por tarea

Específicos para lo que el agente tiene que hacer, nunca generales.

Permisos que vencen

Temporales, con vencimiento al terminar la tarea: acceso justo a tiempo y en la medida justa.

Identidad propia

El agente se autentica con su propia cuenta en lugar de heredar todos los permisos del usuario.

Confirmación y bloqueo

Confirmación explícita para operaciones sensibles; bloqueo automático de lo que esté fuera de su dominio.

SEGÚN EL TIPO DE AGENTE

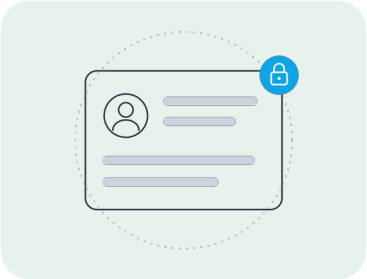
Agentes que usan APIs	Acceso mediado y verificación antes de ejecutar cada acción.
Agentes que operan una computadora	Entornos restringidos y una lista explícita de acciones permitidas.
Orquestadores	Aislamiento entre componentes y límites de seguridad fuertes entre módulos.

EJEMPLO DE LA NORMA, EN NUESTRAS PALABRAS

Un asistente de rendición de gastos puede leer los comprobantes del empleado y crear borradores, pero no entrar a la nómina ni a gastos de otros. Usa su propia cuenta, credenciales que vencen al terminar, y el envío final lo confirma la persona.

04 · CLÁUSULA 6.11.2

Menos datos, menos superficie de ataque



El impacto de un ataque a la confidencialidad o la integridad baja si se reduce la cantidad, la variedad y el tiempo de conservación de los datos. La norma propone cuatro prácticas.

01

Minimización

Dejar afuera del entrenamiento los datos y campos que no aportan a la exactitud, y quitarlos también de los reentrenamientos. El análisis estadístico ayuda a encontrarlos.

02

Retención mínima

Definir cuánto tiempo se guarda cada dato y borrarlo cuando ya no tiene propósito, aun antes de que venza el plazo.

03

Cifrado y ofuscación

Enmascaramiento, cifrado, tokenización y privacidad diferencial, que permite aprender de la población sin exponer a cada individuo.

04

Menos detalles públicos

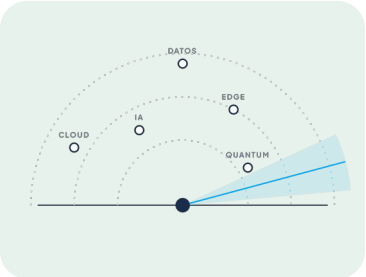
Tratar los detalles técnicos del sistema como un activo: lo que se publica puede ayudar a un atacante a elegir su ataque.

EL EQUILIBRIO

Reducir lo que se publica no debe chocar con la transparencia que exigen las leyes y los clientes. La clave es no regalar lo que ayuda a atacar, sin dejar de explicar cómo funciona el sistema. Para datos personales, la norma remite a ISO/IEC 27701 y 27091.

04 · CLÁUSULA 6.11.3

Gobernar lo que el modelo hace en operación



No todo comportamiento indeseado es un ataque. Antes de reaccionar, conviene averiguar la causa: datos de entrenamiento insuficientes, un modelo desactualizado, errores de ingeniería, un ataque, la sobreconfianza de los usuarios o un sistema con más funciones y permisos de los que debería.

Supervisión	Reglas, guardrails o revisión humana para detectar y frenar decisiones indeseadas. Cuidado: quien solo confirma por costumbre deja de supervisar.
Transparencia	Contar a las partes interesadas cómo funciona el sistema, cómo se entrenó y qué exactitud esperar. Lo mínimo: avisar que hay IA.
Validación continua	Probar el sistema con frecuencia para detectar efectos duraderos de un ataque, como un envenenamiento.
Explicabilidad	Ayuda a confiar y a evaluar riesgos, pero puede revelar qué atributos pesan más y facilitarle el trabajo al atacante.
Pruebas de sesgo	Medir el sesgo con regularidad también sirve para detectar ataques que vuelven discriminatorio al sistema.

REFERENCIAS

La norma remite a ISO/IEC 42001 para la supervisión humana y la gestión de datos, a ISO/IEC TS 6254 para la explicabilidad y a ISO/IEC TS 12791 para el tratamiento del sesgo.

04 · CLÁUSULA 6.11.4

Procedencia: saber de dónde viene todo



La procedencia es la historia documentada de cada activo de IA: su origen, cómo se desarrolló, qué transformaciones sufrió y cómo se usa. Sin ella, no hay forma de detectar una manipulación ni de defender una decisión ante un auditor.

QUÉ REGISTRAR

- Datasets de entrenamiento y de inferencia, arquitectura, dependencias, configuraciones y pipelines de datos
- En IA generativa, también los prompts y los datos de ajuste fino
- Origen de los datos, licencias, preprocesamiento y terceros que intervinieron
- Versiones, despliegues, anomalías y eventos de seguridad, en registros inmutables

CÓMO PROTEGERLA

- Firmar digitalmente los metadatos de procedencia para que nadie los altere
- Verificar la procedencia de todo sistema o dataset de terceros antes de incorporarlo
- Usar formatos estándar cuando sea posible, como los de C2PA o JPEG Trust
- Conservar los registros para auditorías, evaluaciones de cumplimiento y análisis post-incidente

CONEXIÓN CON ISO/IEC 42001

La procedencia de los datos es un control del Anexo A de ISO/IEC 42001 (A.7.5). Un buen registro sirve para las dos normas a la vez.

Aprendizaje federado: ventajas y nuevos riesgos



En el aprendizaje federado, muchos clientes, como dispositivos o sucursales, entrenan el modelo con sus datos locales y un servidor central combina los resultados. Los datos no salen de donde están.

01
El servidor distribuye el modelo global

02
Cada cliente lo entrena con sus datos

03
Los clientes devuelven el modelo actualizado

04
El servidor los combina en un nuevo modelo global

Ventajas

- Cumplimiento: facilita respetar leyes que limitan la transferencia de datos entre países
- Confidencialidad: los datos no salen del cliente
- Menos transferencia de datos y menor costo de ancho de banda

Desafíos de seguridad

- El modelo igual puede revelar datos por inferencia o robo
- Una red enorme de clientes, alguno puede ser malicioso y envenenar el modelo
- Dispositivos muy distintos, con sus propios riesgos de seguridad

A TENER EN CUENTA

El envenenamiento indiscriminado puede llegar a romper la disponibilidad de un sistema federado. Federar no reemplaza los controles: los traslada a cada cliente.

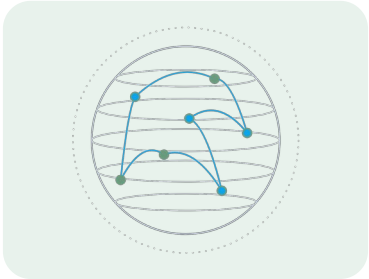
No es otro sistema de gestión



La 27090 se apoya en lo que una organización madura ya tiene. Su lugar natural es la evaluación de riesgos del SGSI y del sistema de gestión de la IA.

ISO/IEC 27001 y 27002 SGSI y controles de seguridad	La 27090 entra por la evaluación de riesgos: ahí se deciden los controles para los sistemas de IA.
ISO/IEC 42001 Sistema de gestión de la IA, certificable	Pide gestionar los riesgos de la IA; la 27090 aporta el detalle de las amenazas técnicas.
ISO/IEC 23894 Guía de gestión de riesgos de IA	Una explica cómo gestionar el riesgo; la otra, cuáles son las amenazas concretas.
ISO/IEC 22989 y 23053 Conceptos y marco de sistemas de IA	Son las dos referencias normativas de la 27090: el vocabulario común.
ISO/IEC 5338 Procesos del ciclo de vida de la IA	Ordena las etapas sobre las que la 27090 ubica cada amenaza.
ETSI EN 304 223, NIST y OWASP Referencias técnicas de seguridad de la IA	La norma remite a ETSI y al SSDF de NIST; el OWASP Top 10 LLM se solapa con la cláusula 6.

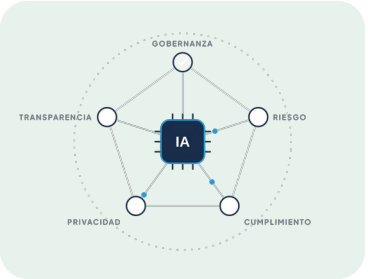
Los controles que la 27090 nombra



La norma conecta sus recomendaciones con controles concretos de ISO/IEC 27002 y del Anexo A de ISO/IEC 42001. Si ya tiene un SGSI o un sistema de gestión de la IA, este es el punto de entrada.

27002 · 8.15 y 8.16	**Registros de eventos y seguimiento de actividades**	Registrar y monitorear el uso del modelo para detectar patrones de ataque.
27002 · 8.27	**Arquitectura segura y principios de ingeniería**	Base de los principios de zero trust aplicados a la IA.
27002 · 8.25 y 8.30	**Ciclo de vida del desarrollo y desarrollo externalizado**	Desarrollo seguro de modelos, propios o de terceros.
27002 · 8.31	**Separación de entornos**	Los datos sensibles suelen estar en desarrollo: protegerlo como producción.
27002 · 5.19 a 5.23	**Proveedores, cadena de suministro TIC y nube**	Evaluar a quienes proveen modelos, datasets y servicios de IA.
42001 · A.6.2.3	**Documentación del diseño y desarrollo**	Considerar las amenazas propias de cada eslabón de la cadena.
42001 · A.6.2.6	**Operación y monitoreo del sistema de IA**	Llevar la detección de la 27090 a la operación diaria.
42001 · A.7.5	**Procedencia de los datos**	El registro de origen y transformaciones que pide la 6.11.4.

El artículo 15 habla el mismo idioma



El artículo 15 del Reglamento (UE) 2024/1689 exige que los sistemas de IA de alto riesgo resistan los intentos de terceros de alterar su uso, sus salidas o su desempeño. Y nombra amenazas concretas que son, casi una por una, las de la 27090.

AMENAZA QUE NOMBRA EL ARTÍCULO 15	CLÁUSULA DE LA 27090
Envenenamiento de datos	6.2
Envenenamiento del modelo	6.7
Ejemplos adversariales o evasión	6.3
Ataques a la confidencialidad	6.4, 6.5 y 6.6
Fallas del modelo	Cláusulas 5 y 7

- Para los sistemas de alto riesgo del anexo III, estas obligaciones se aplican desde el 2 de diciembre de 2027; para los del anexo I, desde el 2 de agosto de 2028.
- La 27090 es voluntaria: aplicarla ayuda a demostrar diligencia, pero no prueba por sí sola el cumplimiento. Lo que corresponde depende del rol (proveedor o usuario) y del nivel de riesgo.

POR QUÉ IMPORTA EN LATINOAMÉRICA

La Ley de IA alcanza a quien pone sistemas en el mercado europeo, esté donde esté. Y aunque su empresa no venda en Europa, sus clientes globales ya preguntan por estas amenazas en sus cuestionarios de seguridad. La 27090 da un idioma común para responder.

Qué escuchamos de los C-levels

> ¿Alguien podría copiar nuestro modelo usando la API?

> ¿Cómo sabemos que nadie alteró los datos de entrenamiento?

> ¿Qué pasa si el asistente obedece una instrucción escondida?

Señales de alerta

- No hay un inventario de los sistemas de IA en uso
- Modelos o datasets descargados sin verificar su origen
- Las APIs de IA no tienen límites de uso ni alertas
- El asistente tiene más permisos de los que necesita
- Nadie prueba el sistema después de reentrenarlo

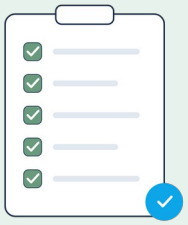
Primeros pasos

- Armar el inventario, incluidos los servicios de terceros
- Sumar las amenazas de la cláusula 6 al análisis de riesgos
- Registrar consultas y respuestas, y llevarlas al SOC
- Aplicar mínimo privilegio a modelos y agentes
- Agendar un primer ejercicio de red teaming

LO QUE PROPONEMOS

Empezar por los sistemas de IA que tocan clientes, dinero o datos personales. Ahí una sola de las nueve amenazas ya justifica el esfuerzo, y la experiencia ganada se replica en el resto.

Cinco preguntas para hacerse antes de poner IA en producción



Si alguna respuesta es «no sé», ahí está el primer paso.

01 **¿Sabemos qué sistemas de IA usamos?**

Incluidos los que vienen dentro de productos de terceros y los que los equipos usan por su cuenta.

02 **¿De dónde vienen los datos y los modelos?**

¿Podemos demostrar su procedencia y que nadie los alteró en el camino?

03 **¿Veríamos un ataque en los registros?**

¿Detectaríamos miles de consultas que intentan copiar el modelo o deducir sus datos?

04 **¿Qué puede tocar el modelo?**

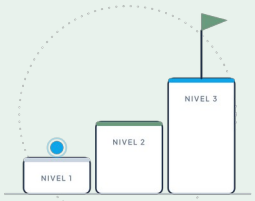
¿A qué datos, sistemas y acciones tiene acceso? ¿Tratamos sus respuestas como no confiables?

05 **¿Lo probamos como lo haría un atacante?**

¿Hacemos red teaming antes de desplegar y después de cada cambio o reentrenamiento?

06 · AUTOEVALUACIÓN

¿Qué tan preparada está su organización?



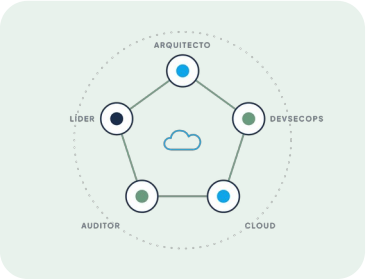
Marque las afirmaciones que hoy se cumplen. Cada una resume una recomendación de la norma.

- [ ] Tenemos un inventario de los sistemas de IA, incluidos los de terceros 01
- [ ] Cada sistema de IA tiene un dueño que decide sobre sus riesgos 02
- [ ] Las amenazas de la cláusula 6 están en nuestro análisis de riesgos 03
- [ ] Conocemos la procedencia de los datos y modelos que usamos 04
- [ ] Verificamos la integridad de los modelos antes de desplegarlos 05
- [ ] Los asistentes y agentes tienen solo los permisos que necesitan 06
- [ ] Registramos entrada, salida, usuario, fecha y versión del modelo 07
- [ ] Limitamos la frecuencia de consultas por usuario 08
- [ ] Tratamos las salidas del modelo como datos no confiables 09
- [ ] Hicimos red teaming antes del despliegue y tras el último cambio 10
- [ ] Revisamos los datos nuevos antes de cada reentrenamiento 11
- [ ] Exigimos evidencia de seguridad a los proveedores de IA 12

CÓMO LEER EL RESULTADO

10 a 12: buena base, a sostener con red teaming periódico. 6 a 9: brechas concretas, priorice los sistemas críticos. Menos de 6: empiece por el inventario y los dueños.

Quién es responsable de qué



La seguridad de la IA no es solo del área de seguridad. Este reparto, adaptado del análisis de la norma, evita que cada amenaza quede «en el medio».

Seguridad de la información	Integra la IA en el análisis de riesgos, define controles y lleva los registros al SOC.
Ingeniería de IA y ML	Aplica desarrollo seguro, pruebas adversariales y controles en el pipeline de entrenamiento.
Gobierno de datos	Asegura la procedencia, la calidad y los permisos de los datos de entrenamiento y de producción.
Compras y proveedores	Exige evidencia de seguridad a quienes proveen modelos, datasets y servicios de IA.
Responsable de gobernanza de IA	Mantiene el inventario, aprueba riesgos residuales y coordina con ISO 42001.
Auditoría interna y cumplimiento	Verifica que los controles existan, funcionen y estén documentados.

CONSEJO

Cada sistema de IA del inventario necesita un dueño que decida ante una amenaza nueva o un cambio en el modelo.

06 · CASO PRÁCTICO

Una fintech con dos sistemas de IA



Un caso ficticio, armado para mostrar cómo se aplica la norma. La empresa tiene un modelo de scoring crediticio que se reentrena cada mes y un asistente con IA generativa que atiende a los clientes y consulta sus cuentas.

Scoring crediticio

6.2 Envenenamiento	Solicitudes falsas aprobadas a propósito para sesgar el reentrenamiento → revisión de los datos nuevos y comparación con el modelo anterior antes de publicarlo.
6.5 Exfiltración	Un competidor podría clonar el modelo con la API → autenticación por cliente, límites de uso y alertas por volumen.

Asistente de clientes

6.9 Inyección de prompts	Un mensaje con instrucciones ocultas → el asistente solo lee datos del cliente autenticado y las transferencias requieren confirmación.
6.8 Salida sensible	Que revele datos de otro cliente → controles de acceso en las fuentes, filtrado de salidas y pruebas de fuga antes de cada versión.

RESULTADO

Los dos sistemas tienen dueño, sus amenazas entran en el análisis de riesgos de ISO 27001 e ISO 42001 y el SOC recibe alertas de uso anómalo.

Los primeros 90 días



Una secuencia realista para una organización que ya usa IA y quiere ordenar su seguridad con la 27090 como referencia.

DÍAS 1 A 15

01

Inventario

Relevar los sistemas de IA propios y de terceros, sus datos, sus modelos y quién responde por cada uno.

DÍAS 16 A 45

02

Evaluación de amenazas

Modelar las amenazas de la cláusula 6 para los sistemas críticos y sumarlas al análisis de riesgos.

DÍAS 46 A 75

03

Controles prioritarios

Mínimo privilegio, límites de uso, filtrado de entradas y salidas, registro y procedencia.

DÍAS 76 A 90

04

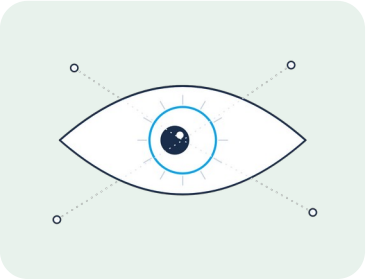
Pruebas y monitoreo

Primer ejercicio de red teaming, alertas en el SOC y criterios de revisión antes de cada reentrenamiento.

DESPUÉS DEL DÍA 90

Repetir el red teaming después de cada cambio importante, revisar el inventario cada trimestre y llevar los resultados a la revisión por la dirección de ISO 27001 e ISO 42001.

Lo que la 27090 no hace



Conocer los límites de la norma evita prometer de más ante la dirección, los clientes o un auditor.

No certifica

Es una guía: no hay certificado ISO/IEC 27090. Para certificar la gestión de la IA, la referencia es ISO/IEC 42001.

No prueba cumplimiento legal

Ayuda a demostrar diligencia, pero no acredita por sí sola el cumplimiento de la Ley de IA europea ni de otra regulación.

No es exhaustiva

La propia norma aclara que sus ejemplos no cubren todo: aparecerán amenazas nuevas a medida que evolucione la IA.

La detección sigue siendo difícil

Muchos ataques usan consultas válidas, y algunas mitigaciones son prácticas emergentes con costos de rendimiento.

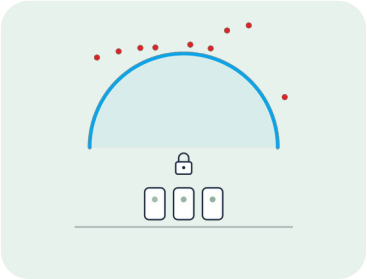
Los agentes no son un tipo de amenaza

Los trata en zero trust (5.3) y mínimo privilegio (6.11.3.3), no en la cláusula 6. Conviene sumar el OWASP Top 10 agéntico.

La privacidad va por otro carril

Para los aspectos de privacidad propios de la IA, la 27090 remite a la ISO/IEC 27091.

De la norma a sistemas de IA protegidos



Primera consultora de Latinoamérica certificada en ISO/IEC 42001. Así llevamos la 27090 a la práctica:

01

Secure AI Lifecycle

Modelado de amenazas para IA, validación de datasets, pruebas de robustez adversarial, Secure MLOps, monitoreo algorítmico y playbooks de incidentes de IA.

02

AI GRC e ISO/IEC 42001

Integramos las amenazas de la 27090 al análisis de riesgos y al sistema de gestión de la IA, listos para auditar.

03

ISO/IEC 27001

Sumamos los sistemas de IA al alcance del SGSI y a la declaración de aplicabilidad.

04

Ley de IA de la UE

Mapeamos el artículo 15 y el resto de las obligaciones según su rol y el nivel de riesgo de cada sistema.

05

Seguridad gestionada

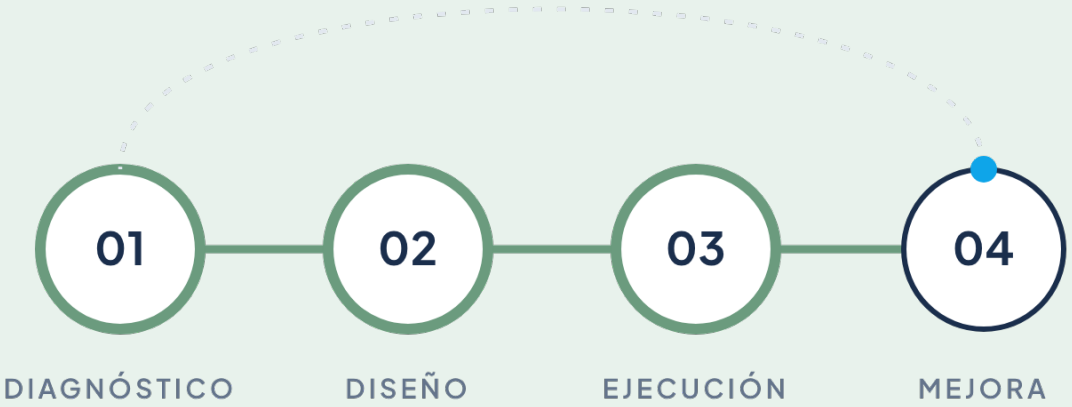
Llevamos los registros de los sistemas de IA al SOC para detectar exfiltración, abuso y comportamientos anómalos.

EMPECEMOS POR UN DIAGNÓSTICO

Relevamos sus sistemas de IA, medimos la brecha contra la 27090 y le entregamos un plan priorizado. Escríbanos a info@cloudlegion.com.ar.

CÓMO TRABAJAMOS

Un método en cuatro etapas



01 **Diagnóstico**
Entendemos tu contexto, tus objetivos y tus riesgos, y medimos la brecha contra la norma o la arquitectura objetivo.

02 **Diseño**
Definimos el plan, los controles y la arquitectura, proporcionales a tu negocio, con responsables y plazos.

03 **Ejecución**
Implementamos junto a tu equipo: documentación, herramientas, capacitación y evidencias.

04 **Mejora continua**
Medimos, auditamos y ajustamos para sostener los resultados y certificar a la primera.

QUIÉNES SOMOS

Arquitectos de la confianza digital

Arquitectos de la confianza digital. Fusionamos la transformación digital con una gobernanza sólida para que la innovación tecnológica, desde la nube hasta la IA, sea el motor de crecimiento seguro de tu organización.

No somos solo consultores, somos legionarios digitales: un cuerpo de élite multidisciplinario que protege el activo más valioso de las organizaciones, su integridad y su capacidad de innovar sin miedo.

+40
legionarios digitales

+500
certificaciones en el equipo

+20
premios y reconocimientos

+100
proyectos en 10 países

100%
de clientes certifica a la primera

+99%
de satisfacción de clientes

NUESTRO CÓDIGO DE HONOR

Las 6 C del Legionario

01

Compromiso

Lealtad absoluta a los objetivos del aliado. Asumimos sus desafíos como propios y co-creamos valor real.

02

Creatividad

Ingenio táctico. Diseñamos soluciones que superan la complejidad del enemigo, con innovación centrada en las personas.

03

Competencia

Liderazgo basado en la investigación y la academia. Nuestra maestría técnica es nuestro escudo.

04

Calidad

Excelencia en cada misión. Hacemos las cosas bien desde la primera vez y medimos nuestro éxito por los resultados de nuestros aliados.

05

Compliance

El cumplimiento es nuestra disciplina. Transformamos el riesgo en control y la norma en una ventaja estratégica.

06

Comunidad

Somos más que una consultora: somos una Legión que comparte conocimiento y hace crecer el ecosistema.

SOMOS REFERENTES

Reconocidos en cada disciplina que ofrecemos

DevOps y DevSecOps

Fundadores de la comunidad DevSecOps Argentina y Latam. Ibiri y Moreira, entre los "50 Influential DevSecOps Professionals". Ambassadors del DevOps Institute y de Wazuh.

Seguridad de la nube

Primera consultora de la región calificada como CSA Global Consultant, con Certified CSA STAR Lead Auditor y líderes del capítulo argentino de la CSA. Microsoft MVP en Azure, Developer Technologies y Security AI.

Cumplimiento normativo

Auditores e implementadores acreditados por PECB, CQI-IRCA y la World Compliance Association. El 100% de nuestros clientes certifica a la primera y el 95% nos vuelve a elegir.

Inteligencia artificial

Primera empresa argentina y primera consultora de Latam certificada en ISO/IEC 42001, con el primer caso de éxito de certificación ISO 42001 de Latinoamérica. Primera consultora de Latam en el AI Trustworthy Pledge de la CSA.

RECONOCIMIENTOS

Más de 20 premios y reconocimientos



- Cybersecurity Awareness Month Champion 2022 y 2023 (NCSA y CISA, EE.UU.)
- Globee Awards 2021, 2023 y 2024: Cybersecurity, CTO of the Year y Business Philanthropist of the Year
- IT World Awards 2020 (Network Products Guide) e Info Security PG Global Excellence 2020
- Cybersecurity Excellence Awards: Best Cybersecurity Startup 2018 y 2019, Team of the Year, Consultant of the Year 2016-2018
- Clutch 2023: Top Cybersecurity Company Argentina y Top Cloud Consulting Company Buenos Aires

PRACTICAMOS LO QUE PROPONEMOS



CLIENTES QUE CONFÍAN EN NOSOTROS

Más de 100 proyectos en 10 países



GLOSARIO

Términos clave



Ataque adversarial o evasión	Entrada modificada a propósito para que el modelo se equivoque.
Datos fuera de distribución (OOD)	Entradas que el entrenamiento no cubrió bien y que pueden provocar errores.
Envenenamiento	Manipulación de los datos o del modelo para alterar lo que aprende.
Exfiltración del modelo	Recrear un modelo a partir de consultas válidas.
Jailbreaking	Inyección de prompts que busca saltar las restricciones del sistema.
Modelo fundacional	Modelo entrenado con muchos datos que se adapta a tareas diversas.
Privacidad diferencial	Técnica que agrega ruido para que ningún individuo influya en el resultado.
Procedencia	Historia documentada de un activo de IA: origen, cambios y uso.
RAG	Generación aumentada por recuperación: el modelo responde con documentos que se agregan al contexto.
Red teaming	Pruebas que emulan a un adversario real para encontrar vulnerabilidades.
Entorno de ejecución confiable (TEE)	Entorno aislado del sistema operativo principal para proteger código y datos.
Zero trust	Nada es confiable por defecto: se verifica cada pedido y se da el mínimo acceso.

FUENTES

Para seguir leyendo

Esta guía se basa en el texto de la norma y en fuentes consultadas en noviembre de 2026. La norma completa se adquiere en ISO o en los organismos nacionales de normalización.

ISO/IEC 27090:2026, Cybersecurity — Artificial intelligence (AI) — Addressing security threats and compromises to AI systems
Primera edición, octubre de 2026

ISO/IEC 27001:2022, ISO/IEC 27002:2022 e ISO/IEC 42001:2023
Normas citadas por la 27090

ISO: ficha de la norma ISO/IEC 27090:2026
iso.org/standard/56581.html

BSI: New guidance to help organizations protect AI from security threats (27/10/2026)

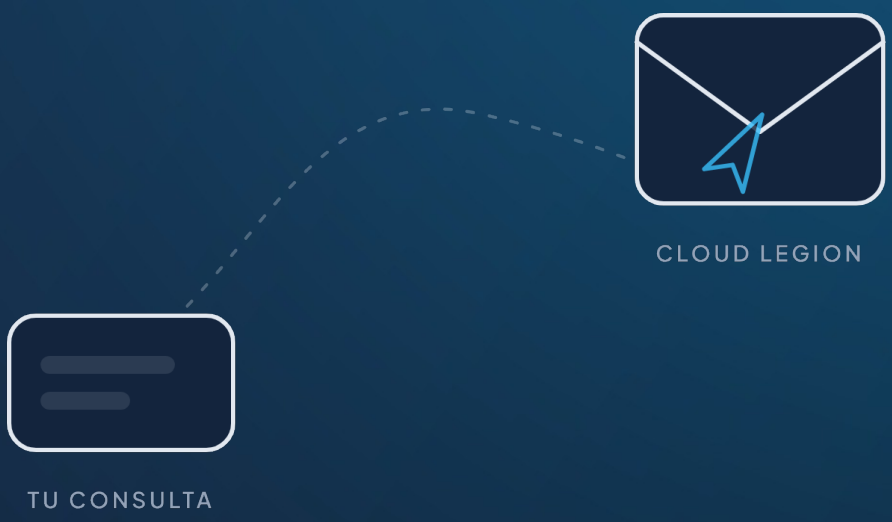
bsigroup.com

Comisión Europea, AI Act Service Desk: artículo 15
ai-act-service-desk.ec.europa.eu/en/ai-act/article-15

AI Act Explorer: artículo 15 y fechas de aplicación
artificialintelligenceact.eu/article/15

OWASP Gen AI Security Project: Top 10 for LLM Applications 2025
genai.owasp.org/llm-top-10

Este ebook es material de divulgación de Cloud Legion. No reproduce el texto de la norma ni reemplaza su lectura. Las referencias a cláusulas siguen la primera edición (2026).



HABLEMOS

¿Listo para tu próximo desafío?

EMAIL	info@cloudlegion.com.ar
WHATSAPP	+54 9 11 4083-3833
OFICINAS	Av. Jujuy 2085, piso 2 "C", Parque Patricios, CABA
WEB	cloudlegion.com.ar
LINKEDIN	linkedin.com/company/cloud-legion