ISO/IEC 27090 explicada

Traducción no oficial adaptada al español de la nota divulgativa publicada por ISO

Fuente original: ISO/IEC 27090 explained

Norma relacionada: ISO/IEC 27090:2026 — ficha oficial y adquisición

** En Perú la NTP-ISO/IEC 42001:2025 es la adopción peruana de la ISO/IEC 42001:2023

¿Por qué esta guía?

¿Cómo se protege un sistema de inteligencia artificial cuando un ataque puede dirigirse no solo contra su software, sino también contra los datos con los que aprende y contra la forma en que se comporta su modelo?

Las organizaciones que desarrollan o utilizan inteligencia artificial necesitan respuestas claras a cuestiones como las siguientes:

- ¿Qué es la ISO/IEC 27090?
- ¿A quién está dirigida?
- ¿Por qué los sistemas de inteligencia artificial necesitan medidas de seguridad adicionales?
- ¿Qué amenazas aborda?
- ¿Cómo se relaciona con la ISO/IEC 27001 y la ISO/IEC 42001?

Esta guía presenta el documento, sus principales ámbitos de orientación y algunos puntos de partida prácticos para entender la seguridad de los sistemas de inteligencia artificial.

1. ¿Qué es la ISO/IEC 27090?

La ISO/IEC 27090 aborda las amenazas de seguridad y las formas de compromiso que afectan específicamente a los sistemas de inteligencia artificial. Su propósito es ayudar a las organizaciones a entender las consecuencias de estas amenazas y los enfoques que pueden utilizar para detectarlas y mitigar sus efectos durante todo el ciclo de vida de un sistema de IA.

La ISO/IEC 27090 está actualmente en fase de publicación. Esta guía se basa en el borrador final; para conocer el contenido definitivo, debería ser consultado con la edición publicada.

Nota del traductor: La ISO/IEC 27090:2026 se publicó en octubre de 2026. La edición oficial puede consultarse y adquirirse en el catálogo de ISO.

Su enfoque es **proteger los sistemas de inteligencia artificial frente a amenazas de seguridad**, y no utilizar la inteligencia artificial para defender otros sistemas. Proporciona orientación en materia de seguridad, no requisitos para un sistema de gestión de inteligencia artificial.

Nota del traductor: La ISO/IEC 27090 es un documento de orientación técnica. Sus recomendaciones ayudan a comprender las amenazas específicas para la IA y a considerar posibles medidas de protección, pero no tienen que interpretarse como requisitos de un sistema de gestión de IA ni como criterios independientes de certificación. Esto no significa que las organizaciones no requieran cumplir otras obligaciones legales, contractuales o normativas que les resulten aplicables.

Esta distinción es importante al relacionarla con la ISO/IEC 42001, que establece requisitos para un sistema de gestión de inteligencia artificial.

2. ¿A quién está dirigida la ISO/IEC 27090?

El documento está dirigido a **organizaciones de cualquier tamaño que desarrollen o utilicen inteligencia artificial**. Esto incluye empresas, entidades públicas y organizaciones sin fines de lucro.

Puede servir como referencia para facilitar el diálogo entre los equipos de ciberseguridad, los desarrolladores de IA, los responsables de los datos y quienes supervisan la adquisición o el uso de sistemas de inteligencia artificial.

Aunque estos grupos tienen responsabilidades diferentes, necesitan compartir un entendimiento de los elementos que podrían verse comprometidos y de las consecuencias que ello tendría.

El alcance del documento no se limita a la inteligencia artificial generativa ni a los chatbots. Aborda las amenazas de seguridad que pueden afectar a los sistemas de IA en un sentido más amplio.

Nota del traductor: En este contexto, las partes interesadas comprenden a las personas, los grupos de personas y las organizaciones que pueden afectar, verse afectados por una decisión o actividad, o considerar que dicha decisión o actividad afecta a sus intereses. Pueden incluir a la dirección, los usuarios, los clientes, los proveedores tecnológicos y las personas afectadas por las decisiones de un sistema de IA.

3. ¿Por qué los sistemas de inteligencia artificial necesitan medidas de seguridad adicionales?

Los sistemas de inteligencia artificial siguen expuestos a las amenazas convencionales de ciberseguridad. Por ello, continúa siendo necesario proteger el software, la infraestructura tecnológica y los mecanismos de acceso a la información. Las medidas específicas para la IA complementan esta base de protección; no la sustituyen.

Muchos sistemas de IA también dependen considerablemente de sus datos de entrenamiento. Si un atacante introduce datos maliciosos, sus efectos pueden trasladarse al comportamiento del modelo. Asimismo, los datos recopilados durante el funcionamiento del sistema pueden generar una exposición adicional cuando se utilizan posteriormente para el re-entrenamiento

Nota del traductor — Datos de entrenamiento: Son los datos que se utilizan para que un modelo de IA aprenda patrones y relaciones que después empleará para producir resultados. Si estos datos se manipulan deliberadamente, el modelo puede aprender patrones no deseados o generar resultados incorrectos.

Otros ataques pueden explotar accesos que parecen legítimos. Por ejemplo, un atacante puede realizar consultas repetidas a un modelo con la intención de extraerlo. Por esta razón, controlar quién tiene acceso al sistema puede resultar insuficiente si no existen mecanismos para detectar usos sospechosos.

Nota del traductor — Extracción de un modelo: Se refiere a intentos de obtener un modelo o reconstruir parte de su funcionamiento mediante las respuestas que proporciona u otras formas de acceso. Puede tener como finalidad apropiarse de propiedad intelectual o facilitar ataques posteriores.

Las consecuencias pueden extenderse más allá de la pérdida de información o de propiedad intelectual. Cuando la IA interviene en la toma de decisiones o se utiliza en entornos donde la seguridad de las personas es crítica, la alteración de su comportamiento también puede causar daños a las personas.

4. ¿Qué aborda la ISO/IEC 27090?

El documento reúne tres ámbitos interconectados: la aplicación de la seguridad de la información a la IA, entender los potenciales compromisos y considerar como las medidas de salvaguarda trabajan en conjunto.

Aplicación de la seguridad de la información a la inteligencia artificial

Entre los temas abordados se encuentran el gobierno de la inteligencia artificial, los principios de confianza cero, las prácticas de ingeniería de software y la seguridad de la cadena de suministro.

El documento también contempla el modelado de amenazas y el red teaming: enfoques que permiten examinar cómo podría ser atacado un sistema y someterlo a pruebas desde una perspectiva adversarial.

Notas del traductor:

- **Gobierno de la inteligencia artificial:** comprende las estructuras de autoridad, las responsabilidades, los criterios de decisión y los mecanismos de supervisión mediante los cuales una organización dirige y controla el desarrollo y uso de la IA.
- **Confianza cero (Zero Trust):** enfoque de seguridad que evita conceder confianza implícita a usuarios, dispositivos o aplicaciones por el solo hecho de encontrarse dentro de una red o haber sido autorizados anteriormente. El acceso tendría que verificarse según corresponda.
- **Seguridad de la cadena de suministro:** considera los riesgos relacionados con terceros, proveedores, componentes de software, conjuntos de datos, modelos y otros elementos externos de los que depende un sistema de IA.
- **Modelado de amenazas:** proceso sistemático para identificar posibles atacantes, objetivos, vulnerabilidades y escenarios de ataque, con el fin de orientar las medidas de protección.
- **Red teaming:** ejercicio en el que un equipo adopta la perspectiva de un atacante para poner a prueba las defensas de un sistema. En IA puede utilizarse para explorar vulnerabilidades y evaluar la eficacia de las medidas de protección.

Entendiendo las amenazas específicas para la inteligencia artificial

El documento examina diferentes tipos de compromiso de seguridad, sus posibles consecuencias y los mecanismos para detectarlos y mitigar sus efectos. Entre ellos se incluyen los siguientes:

- **Envenenamiento de datos (*data poisoning*):** introducción de datos maliciosos o no deseados en los datos de entrenamiento para influir en los resultados del modelo.

Nota del traductor: El ataque consiste en manipular deliberadamente los datos para alterar lo que el modelo aprende o las respuestas que produce.

- **Extracción del modelo (*model exfiltration*):** obtención de un modelo o de información que permite reconstruirlo o reproducir parte de su comportamiento, incluso mediante consultas que pueden parecer legítimas.

- Otras amenazas relacionadas con el comportamiento del modelo o la exposición de información, entre ellas los ataques de evasión, la inferencia de pertenencia, la inversión del modelo y el envenenamiento directo del modelo.

> **Nota del traductor:** Estos conceptos describen técnicas diferentes:
>
> - **Ataques de evasión (evasion attacks):** buscan que el sistema produzca un resultado incorrecto o no deseado mediante entradas diseñadas para eludir su comportamiento previsto.
> - **Inferencia de pertenencia (membership inference):** intenta determinar si determinados datos formaron parte del conjunto utilizado para entrenar un modelo. Esto puede revelar información sensible sobre los datos de entrenamiento.
> - **Inversión del modelo (model inversion):** intenta inferir o reconstruir información sobre los datos utilizados para entrenar un modelo a partir de sus respuestas u otras características observables.
> - **Envenenamiento directo del modelo (direct model poisoning):** busca alterar maliciosamente el propio modelo o sus parámetros, en lugar de limitarse a manipular los datos que recibe durante su funcionamiento.

- Amenazas relacionadas con las entradas y salidas del sistema, incluidas la inyección de instrucciones (*prompt injection*), las salidas del modelo que contienen información sensible y las respuestas que incorporan contenido diseñado para desencadenar ataques.

> **Nota del traductor — Inyección de instrucciones:** Ocurre cuando se introducen indicaciones maliciosas en los datos o contenidos que procesa un sistema de IA para intentar que ignore sus instrucciones previstas, revele información o realice acciones no autorizadas. Por ejemplo, un documento consultado por un asistente de IA podría contener instrucciones ocultas o explícitas destinadas a manipular su comportamiento.

El documento también aborda el origen y el historial de los sistemas de IA y de sus datos, así como las formas de limitar los efectos de comportamientos no deseados del modelo durante su funcionamiento.

> **Nota del traductor:** Conocer el origen y la procedencia de los datos y modelos ayuda a investigar incidentes, identificar posibles manipulaciones y comprender cómo se incorporaron determinados comportamientos al sistema. Esta información también puede facilitar la determinación de responsabilidades y la adopción de medidas correctivas.

Mantener la protección a lo largo del tiempo

Una medida de protección no se evalúa únicamente de manera aislada. El documento examina las interacciones entre las medidas de mitigación, incluidos los posibles conflictos entre ellas, y la necesidad de mantener la protección durante todo el ciclo de vida del sistema de inteligencia artificial.

Entre los temas abordados figuran el registro de eventos y la monitorización, la protección de los entornos de desarrollo, la detección de entradas maliciosas y la limitación del uso de los modelos. También se considera la posibilidad de que los modelos y las medidas de mitigación pierdan eficacia con el tiempo.

Notas del traductor:

- **Registro de eventos (logging):** recopilación de registros sobre actividades, operaciones, accesos y eventos relevantes del sistema. Estos registros pueden utilizarse para investigar incidentes.
- **Monitorización:** observación continua o periódica del comportamiento del sistema para identificar anomalías o indicios de actividad maliciosa.
- **Entorno de desarrollo:** conjunto de herramientas, recursos, datos y servicios utilizados para crear, entrenar, probar o modificar un sistema de IA.
- **Pérdida de eficacia de las medidas de mitigación:** una protección inicialmente útil puede dejar de ser suficiente cuando cambian las amenazas, los datos, el modelo o las formas de interacción con el sistema.

El enfoque subyacente es la defensa en profundidad: emplear varias capas de protección, reconociendo al mismo tiempo que la detección de algunas amenazas contra la inteligencia artificial sigue siendo difícil.

Nota del traductor — Defensa en profundidad: Consiste en combinar medidas preventivas, detectivas y correctivas para reducir la probabilidad de que un ataque tenga éxito y limitar sus consecuencias. La seguridad no depende de una única barrera de protección.

5. ¿Cómo se relaciona la ISO/IEC 27090 con otras normas?

ISO/IEC 27001 e ISO/IEC 27002: la base de la seguridad de la información

La ISO/IEC 27001 establece requisitos para un sistema de gestión de la seguridad de la información. La ISO/IEC 27002 proporciona orientación sobre controles de seguridad de la información.

La introducción de la ISO/IEC 27090 recomienda aprovechar las prácticas establecidas en ambas normas. La orientación específica para la IA amplía el análisis de las amenazas y los activos que pueden requerir atención adicional.

Nota del traductor: Un activo es cualquier elemento que tenga valor para la organización y que necesita protegerse, como la información, los datos de entrenamiento, los modelos, el software, la infraestructura o la propiedad intelectual.

La relación entre estas normas puede entenderse de la siguiente manera: la ISO/IEC 27001 aporta los requisitos del

sistema de gestión de la seguridad de la información; la ISO/IEC 27002 proporciona orientación para seleccionar y aplicar controles; y la ISO/IEC 27090 profundiza en las amenazas particulares de los sistemas de inteligencia artificial y en las medidas que pueden considerarse para abordarlas.

ISO/IEC 42001: Gestión de la inteligencia artificial en toda la organización

La ISO/IEC 42001 establece requisitos para un sistema de gestión de inteligencia artificial. Aborda la forma en que una organización gestiona sus responsabilidades, riesgos y oportunidades relacionados con la IA.

La diferencia principal está en el enfoque: la ISO/IEC 42001 se ocupa de la gestión organizacional de la inteligencia artificial, mientras que la ISO/IEC 27090 examina las amenazas de seguridad que afectan a los sistemas de IA y las formas de abordarlas.

La seguridad es una parte importante de la gestión responsable de la inteligencia artificial, pero no abarca por sí sola todas las cuestiones a considerarse.

Nota del traductor — Cómo se complementan ambas normas: La ISO/IEC 42001 permite establecer un sistema de gestión para dirigir y controlar las actividades relacionadas con la IA. La ISO/IEC 27090 puede utilizarse como fuente de orientación especializada para profundizar en la seguridad de los sistemas de IA.

En términos prácticos, una organización puede utilizar la ISO/IEC 42001 para organizar sus responsabilidades y procesos de gestión, y recurrir a la ISO/IEC 27090 para comprender mejor las amenazas específicas que pueden considerarse al desarrollar, adquirir o utilizar soluciones de IA.

ISO/IEC 22989 e ISO/IEC 23053: conceptos compartidos

El documento cita las normas ISO/IEC 22989 e ISO/IEC 23053:2022 como referencias normativas. Estas proporcionan, respectivamente, conceptos y terminología sobre IA, así como un marco para describir sistemas de IA que emplean aprendizaje automático. Estos fundamentos ayudan a los lectores a comprender los sistemas y conceptos tratados en la guía de seguridad.

Nota del traductor: En este contexto, una referencia normativa es un documento citado por otro documento y necesario para aplicar o comprender determinadas partes de su contenido. Esta denominación no significa que todas las referencias citadas sean, por sí mismas, normas que establezcan requisitos para una certificación.

6. Primeros pasos prácticos

Las organizaciones que se preparen para utilizar las orientaciones de la ISO/IEC 27090 pueden comenzar por las siguientes actividades:

- **Identificar los sistemas de IA que desarrollan o utilizan**, incluidos aquellos proporcionados por terceros.
- **Comprender el origen de sus datos y modelos**, así como los cambios que experimentan durante el funcionamiento, la actualización o el re-entrenamiento.
- **Reunir a los responsables de ciberseguridad, desarrollo de IA, gestión de datos y adquisiciones**, para construir una visión compartida de las amenazas y las responsabilidades.
- **Revisar las prácticas de seguridad existentes**, con el fin de identificar los ámbitos en los que las amenazas específicas de la IA requieren un análisis más detallado.
- **Determinar cómo se detectarían los comportamientos sospechosos y quién tendría que responder**, considerando los mecanismos de comunicación, análisis y tratamiento de incidentes.

Estas actividades constituyen puntos de partida para el análisis y la discusión. No sustituyen la evaluación de un sistema en particular ni la consulta de la edición oficial publicada del documento.

Nota del traductor — Aplicación según el contexto: No todos los sistemas de IA requieren las mismas medidas de seguridad. La profundidad del análisis y las salvaguardas que se consideren guardan relación con el contexto de uso, la naturaleza de los datos, las capacidades del sistema, las amenazas identificadas y las posibles consecuencias de un compromiso de seguridad.

La ISO/IEC 27090 puede ayudar a las organizaciones a formular mejores preguntas, reconocer amenazas que podrían pasar inadvertidas y fundamentar sus decisiones de protección. Su valor reside en aportar orientación especializada que complemente las prácticas de seguridad de la información y la gestión organizacional de la inteligencia artificial.

Fuentes

- **Nota divulgativa de ISO:** ISO/IEC 27090 explained
- **Norma publicada:** ISO/IEC 27090:2026 — ficha oficial para consulta y adquisición

Nota final: La presente publicación corresponde a una traducción adaptada de la nota divulgativa de ISO, con aclaraciones editoriales y notas del traductor para facilitar su comprensión. No es una traducción oficial de ISO ni reproduce íntegramente la norma ISO/IEC 27090:2026. Para conocer el contenido técnico completo y definitivo, se recomienda consultar y adquirir la edición oficial.