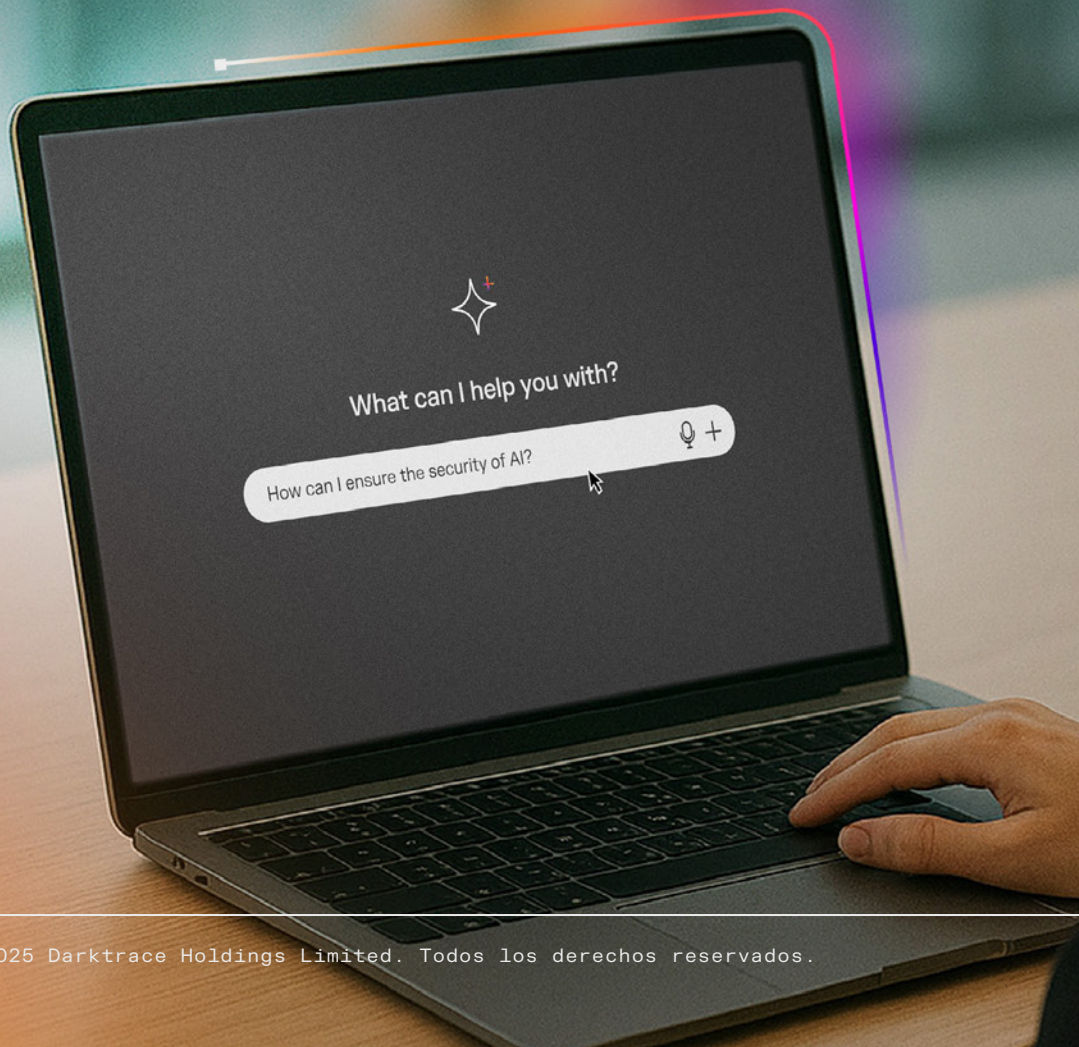


Cómo proteger la IA en la empresa

Una guía para proteger los modelos, los datos y los agentes



Contenido

02	Prefacio	07	Escaneo de laC
02	Acerca de Darktrace	07	Escaneo de código generado por IA
03	La seguridad de la IA: comprender qué está en riesgo	08	Seguridad de la cadena de suministro de IA
04	5 categorías de riesgo de la IA y cómo abordarlas	08	Uso de la IA en la sombra
05	Prevención del uso indebido y de los comportamientos emergentes	08	Proveedores de IA que proporcionan agentes directamente
05	Análisis de prompts maliciosos	08	La IA en los servicios a proveedores
05	Casos de uso no intencionados	08	Escaneo de código y dependencias de agentes de IA
05	Acceso al historial de prompts	09	Suministro de modelos de IA
05	Efectos de resultados maliciosos y alucinatorios	09	Datos de entrenamiento de IA
06	Supervisión y control de la IA en funcionamiento	10	Fortalecimiento de la preparación y la supervisión
06	Acciones y conexiones de los agentes de IA	10	Informes
06	Gateway de IA	10	Pruebas de penetración con IA, equipos rojos y morados
06	Captura de estado del agente de IA	10	Formación y simulacros del equipo de seguridad
06	Transmisión de datos de IA	11	Un nuevo enfoque de la seguridad de la IA para los consejos de administración
07	Protección del desarrollo e infraestructura de la IA	12	Conclusión
07	Configuración de la seguridad de la IA		
07	Arquitecturas de agentes de IA		

Prefacio

Hemos escuchado a nuestros clientes y observado de primera mano que los líderes de seguridad carecen de una definición clara y concisa de lo que implica «proteger la IA». Aunque existen recursos valiosos, aún no hemos encontrado uno que alinee de forma cohesionada el riesgo con las funciones de seguridad. Los marcos más amplios, como la ISO/IEC 42001, proporcionan importantes principios de gobernanza, pero no ofrecen la orientación práctica necesaria para hacer operativa la seguridad de la IA en una empresa.

El presente documento tiene por objeto subsanar esa carencia. Ofrece una guía independiente de proveedores para comprender el alcance de la ciberseguridad en el ámbito de la IA, en la que se describen las categorías clave, los riesgos y las funciones de seguridad que las organizaciones deben tener en cuenta a la hora de adoptar e integrar tecnologías de IA en sus actividades empresariales.

Acerca de Darktrace

Darktrace lleva más de una década a la vanguardia de la aplicación de la inteligencia artificial a la ciberseguridad y es pionera en innovaciones que han transformado el sector. En 2013, lanzamos la solución de detección y respuesta en red (NDR), basada íntegramente en métodos de detección impulsados por la IA, lo que estableció un nuevo estándar para la identificación autónoma de amenazas.

En 2019, presentamos un agente de investigación autónomo basado en IA que reproduce la investigación y el análisis humanos a la velocidad y a la escala propias de una máquina. Nuestra temprana adopción de modelos lingüísticos avanzados, entre los que se incluyen los enfoques basados en transformers en 2021 y los modelos lingüísticos específicos de dominio (DSLML) en 2023, reforzó aún más nuestra capacidad para interpretar ciberamenazas complejas.

En la actualidad, Darktrace AI opera de forma continua en aproximadamente 10 000 organizaciones de todo el mundo y ofrece una resiliencia contrastada a gran escala. A mediados de 2025, nos convertimos en una de las primeras empresas de ciberseguridad en obtener la certificación ISO 42001, lo que subraya nuestro compromiso con una gobernanza responsable de la IA. Todo ello refleja no solo nuestro liderazgo en innovación en materia de IA, sino también nuestra inquebrantable dedicación a garantizar la seguridad de la propia IA, asegurándonos de que la confianza, la transparencia y la seguridad sigan siendo el núcleo de cada avance.



Más de 10 años de aplicación de la IA a casos de uso en el ámbito de la ciberseguridad



Una de las primeras empresas de ciberseguridad en obtener la acreditación ISO 42001



Ganadora del AI Excellence Awards en IA de Ciberseguridad (2025)



Alrededor de 10 000 clientes y la cifra sigue aumentando

La seguridad de la IA

Comprender qué está en riesgo

La inteligencia artificial se está implantando en las empresas a un ritmo que pocos sistemas de gobernanza y gestión pueden igualar.

Los marcos tradicionales tienen dificultades para mantenerse al día, ya que la IA introduce nuevas complejidades y amplía la superficie de ataque. Cada modelo, conjunto de datos y agente, junto con la compleja red de API, integraciones de terceros e interacciones humanas que los sustentan, presenta nuevos riesgos y crea nuevas vías de vulnerabilidad.

La IA promete mejoras en productividad, velocidad y eficiencia para algunos, pero también conlleva su propias incertidumbres y una creciente necesidad de supervisión, especialmente en el ámbito de la IA generativa y los agentes basados en GenAI. La IA generativa, que puede producir resultados poco fiables o «alucinaciones», carece de transparencia en la toma de decisiones y opera de forma autónoma en aspectos que las herramientas de seguridad tradicionales nunca fueron diseñadas para supervisar, lo que amplía el ecosistema digital en el que la propiedad, el control y la responsabilidad a menudo se difuminan.

Para los responsables de seguridad, el primer reto a la hora de proteger una empresa en tiempos de transformación consiste en identificar qué es lo que hay que proteger. La IA está entrando en las organizaciones a través de numerosos canales. Los empleados utilizan ahora herramientas SaaS y de productividad basadas en IA. Los proveedores de servicios en la nube y locales están ampliando su infraestructura y sus servicios con capacidades de IA integradas. Los proveedores y el software adquirido suelen introducir funciones de IA, lo que aumenta la exposición sin una supervisión clara. Dentro de la propia empresa, los desarrolladores pueden crear o adaptar modelos y agentes propios que integren la IA directamente en el negocio.

En todos estos puntos de acceso, la protección de un ecosistema de IA requiere visibilidad en todas las capas: desde el desarrollo y el entrenamiento hasta la implementación y el funcionamiento en tiempo real. Cada etapa presenta unos riesgos propios, que van desde la manipulación maliciosa de datos y los fallos de dependencia hasta el comportamiento malintencionado de agentes o el uso no deseado de los resultados de los modelos.

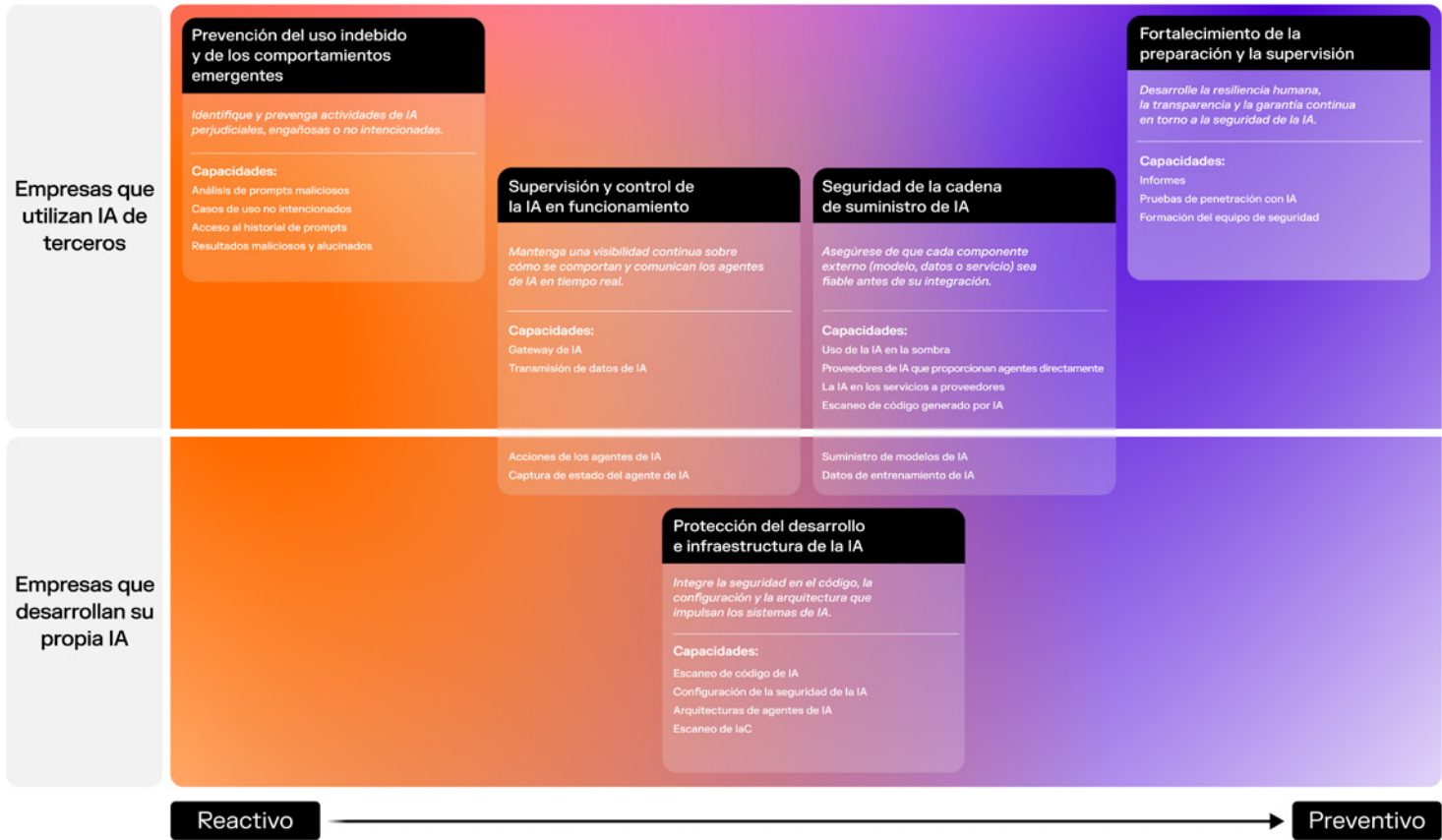


5 categorías de riesgo de la IA y cómo abordarlas

Cada una de las cinco categorías descritas en este marco representa un nivel distinto de seguridad de la IA, desde los proveedores externos y los procesos de desarrollo hasta las operaciones en tiempo real y la gobernanza organizativa.

Dentro de cada una, las capacidades específicas definen dónde puede surgir el riesgo y dónde se pueden aplicar las funciones de seguridad para reducir la exposición. En la siguiente sección se analizan estas capacidades en detalle, relacionando cada una de ellas con los riesgos asociados y las funciones de seguridad diseñadas para hacerles frente. Hemos identificado cinco categorías críticas de uso de la IA. Estas categorías reflejan las formas en que la IA se integra y funciona dentro de una organización: a través de los proveedores, los procesos de desarrollo, los sistemas operativos, las interacciones con los usuarios y los procesos de gobernanza.

En resumen, las cinco categorías son:



La tabla anterior ofrece un resumen general de estas áreas, en el que se describe cómo se relacionan los riesgos con las funciones clave de seguridad a lo largo del ciclo de vida de la IA. Ofrece una visión concisa de dónde deben centrarse las defensas antes de que las siguientes páginas exploren cada capacidad con mayor profundidad.

Prevención del uso indebido y de los comportamientos emergentes

Capacidades

Análisis de prompts maliciosos

La detección y el análisis de prompts perjudiciales o manipuladores que intentan coaccionar a los sistemas o agentes de IA para que realicen acciones no deseadas o peligrosas.

Descripción del riesgo: se puede ordenar a los agentes de IA —o engañarlos— para que realicen acciones no deseadas utilizando sus permisos y su acceso a los datos. Esta acción puede resultar difícil de supervisar o bloquear, sobre todo en el caso de los agentes de IA externos. Si dispone de acceso suficiente, un actor de amenazas puede explorar y extraer datos de un modelo de IA que estaban destinados a permanecer ocultos en su interior.

Función de seguridad: detecte y, en la medida de lo posible, bloquee las solicitudes con intenciones maliciosas, ya sean internas o externas, directas o indirectas. Supervise los intentos de extracción o manipulación de los datos del modelo para evitar usos indebidos y proteger la información confidencial en el marco de las interacciones de IA.

Casos de uso no intencionados

El uso indebido o la reutilización de los sistemas de IA para fines distintos a los previstos o autorizados inicialmente, lo que a menudo da lugar a comportamientos o riesgos imprevistos.

Descripción del riesgo: los agentes y sistemas de IA que se utilizan de forma indebida pueden tener una serie de efectos indeseados. Estos riesgos se ven agravados por la dificultad de detectar el uso indebido por parte de personas autorizadas, en caso de que las medidas de control no sean eficaces.

Función de seguridad: identifique y bloquee los usos no autorizados o no aprobados de los sistemas de IA. Supervise la exposición involuntaria de datos o el uso indebido de los mismos para garantizar que los modelos de IA funcionen únicamente dentro de los parámetros aprobados y conformes a la normativa.

Acceso al historial de prompts

El almacenamiento y la gestión de los historiales de prompts de IA, que pueden contener datos confidenciales, información contextual o patrones de comportamiento.

Descripción del riesgo: el acceso no autorizado al historial de prompts puede revelar información confidencial, aunque el servicio al que se enviaron los prompts sea seguro y no conserve dicha información.

Función de seguridad: restrinja el acceso y supervise si se ha producido alguna vulneración, ya que un actor de amenazas que tenga control sobre la cuenta podría aprovechar esta información para llevar a cabo nuevos ataques.

Efectos de resultados maliciosos y alucinatorios

La generación de resultados de IA que sean inexactos, engañosos o intencionadamente perjudiciales, y que puedan influir en las acciones humanas.

Descripción del riesgo: al margen de los riesgos relacionados con la automatización que utiliza resultados erróneos o maliciosos (como la apropiación indebida de paquetes y las acciones inesperadas de los agentes de IA), estos resultados también pueden provocar que las personas realicen acciones no deseadas. El alcance de estas posibles acciones es mucho más amplio que el de las acciones automatizadas; por ejemplo, si se les ordena transferir dinero o aplicar configuraciones maliciosas a un sistema informático crítico cuando estos están deliberadamente fuera del alcance de los agentes de IA.

Función de seguridad: supervise los resultados de la IA que sean perjudiciales o falsos y que puedan inducir a error a los usuarios o provocar acciones no deseadas. Implemente la detección de alucinaciones y la validación de los resultados para evitar la desinformación o comportamientos peligrosos influidos por contenidos generados por la IA.

Supervisión y control de la IA en funcionamiento

Capacidades

Acciones y conexiones de los agentes de IA

Para empresas que desarrollan herramientas de IA

La supervisión del comportamiento en tiempo real de los agentes de IA, incluidas las conexiones salientes, los intercambios de datos y las interacciones con otros agentes o servicios.

Descripción del riesgo: los agentes de IA pueden hacer un uso indebido de sus permisos o acceder a datos o acciones que no estaban previstos. También pueden filtrar datos o provocar acciones no deseadas al llamar a otros agentes de IA o a servicios (incluidos los servidores MCP). Los agentes de IA carecen de la cautela del personal humano si las medidas de protección no prevén posibles problemas o si fallan.

Función de seguridad: supervise los agentes de IA en tiempo real para detectar conexiones salientes, solicitudes y prompts dirigidos a servicios externos (incluidos los servidores MCP) u otros agentes. Realice un seguimiento del uso de credenciales y aplique bloqueos cuando sea necesario para evitar fugas de datos o acciones no autorizadas.

Gateway de IA

Un punto de acceso centralizado o una capa de control que gestiona las interacciones de los sistemas de IA, como el envío de prompts, el enrutamiento de modelos, el control de costes y el cumplimiento de la normativa sobre privacidad de datos.

Descripción del riesgo: es posible que no se pueda acceder a los sistemas de IA a través de los gateways, o que un gateway esté mal configurado o falle. Ello puede dar lugar a una falta de seguridad en los prompts, un enrutamiento incorrecto de los modelos, costes incontrolados o una supervisión insuficiente, así como a una falta de control en tiempo real.

Función de seguridad: implemente, gestione de forma centralizada y garantice el uso de los gateways de IA siempre que sea posible. Incluya registros, alertas y métricas en los sistemas y en la supervisión de la seguridad.

Cuando se implementan deliberadamente sistemas de IA para conservar los datos a nivel local (por ejemplo, en las propias instalaciones), las capacidades de seguridad deben adaptarse a ello y estar a la altura.

Captura de estado del agente de IA

Para empresas que desarrollan herramientas de IA

El proceso de conservar el estado operativo y la información de los componentes de IA que sean transitorios o efímeros, en particular en entornos en la nube, con el fin de facilitar la investigación y el análisis posteriores a un incidente.

Descripción del riesgo: si un componente efímero de un agente de IA (algo habitual en los sistemas alojados en la Nube) realiza alguna acción que requiera ser investigada o conservada, dicho componente podría desaparecer antes de que los procesos de seguridad puedan capturarlo. Es posible que el equipo de seguridad no disponga de la capacidad necesaria para capturar y conservar determinados componentes.

Función de seguridad: capture y conserve de forma automática el estado completo de un agente de IA sospechoso, incluso las instancias transitorias en la nube. Así podrá llevar a cabo investigaciones y tomar medidas, al tiempo que se garantiza la conservación de las pruebas y el contexto para su análisis.

Las industrias altamente reguladas deben prestar especial atención a este riesgo

Transmisión de datos de IA

La transferencia de datos entre componentes de IA, sistemas o endpoints externos, incluidos los intercambios tanto automatizados como iniciados por personas.

Descripción del riesgo: el envío de datos a sistemas de IA puede vulnerar la soberanía de los datos o los requisitos legales (como el RGPD). Los datos los pueden enviar personas, sistemas automatizados o decisiones tomadas por agentes de IA. Las transmisiones de datos hacia o desde la IA pueden utilizar criptografía débil y deben tenerse en cuenta en los planes de transición hacia la criptografía postcuántica (PQC).

Función de seguridad: proteja los datos en tránsito entre los componentes de IA mediante métodos de cifrado, incluida la criptografía postcuántica cuando proceda. Asegúrese de que todas las transferencias de datos cumplan con la normativa pertinente en materia de circulación de datos y privacidad, como el RGPD.

Protección del desarrollo e infraestructura de la IA

Capacidades

Configuración de la seguridad de la IA

Para empresas que desarrollan herramientas de IA

La configuración y la gestión constante de los parámetros de seguridad, las medidas de protección y los controles de acceso que regulan el funcionamiento de los componentes de IA.

Descripción del riesgo: las configuraciones asignadas a los agentes de IA y a los componentes individuales relacionados y de apoyo pueden presentar fallos y permitir un acceso indebido o adolecer de una seguridad insuficiente frente a los ataques. Una configuración incorrecta también puede eliminar las medidas de protección previstas.

Función de seguridad: garantice configuraciones seguras en todos los componentes de IA, incluidos los mecanismos de control, las políticas y los permisos. La supervisión y el análisis automatizados, junto con revisiones periódicas de las políticas y su aplicación, contribuyen a mantener un nivel de seguridad adecuado y conforme a la normativa en el ámbito de la IA.

Arquitecturas de agentes de IA

Para empresas que desarrollan herramientas de IA

El diseño y la estructura de los agentes de IA y sus componentes interconectados, que determinan cómo interactúan, comparten datos y hacen cumplir los límites de seguridad.

Descripción del riesgo: la arquitectura multicomponente de un agente de IA en uso, o de varios agentes de IA que funcionan conjuntamente en un diseño, puede presentar fallos en su conjunto y permitir un acceso indebido o carecer de medidas de protección contra un uso no deseado, al margen de los problemas que puedan presentar los componentes individuales.

Función de seguridad: realice análisis estáticos de las arquitecturas de los agentes de IA para identificar brechas de seguridad. Evalúe la exposición de la red a nivel local, las credenciales de los servicios y los permisos para evitar el acceso no autorizado o el movimiento lateral. Una segmentación adecuada y una gestión correcta de las credenciales son fundamentales para reducir el riesgo arquitectónico.

Los equipos de desarrollo pueden recurrir a servicios de apoyo para la creación de agentes de IA. Se recomienda racionalizar, en la medida de lo posible, la elección a un único servicio que ofrezca unas capacidades de seguridad aceptables.

Escaneo de IaC

Para empresas que desarrollan herramientas de IA

El análisis automatizado de plantillas de infraestructura como código (IaC) para detectar errores de configuración o vulnerabilidades de seguridad antes de la implementación.

Descripción del riesgo: cuando se utiliza la infraestructura como código, pueden surgir problemas relacionados con la configuración de seguridad de la IA y la arquitectura de los agentes de IA debido a un acceso inadecuado al código, a la introducción de código malicioso, ya sea de forma interna o procedente de una fuente externa, o a errores.

Función de seguridad: Utilice herramientas de análisis especializadas para examinar la infraestructura como código (IaC) en busca de configuraciones incorrectas o riesgos implícitos. La detección temprana de problemas de seguridad en los scripts de implementación evita que las vulnerabilidades se propaguen por los entornos de IA.

Escaneo de código generado por IA

Para empresas que desarrollan herramientas de IA

La revisión del código generado por sistemas de IA (copilotos de codificación) con el fin de detectar fallos de seguridad, inclusiones maliciosas o dependencias poco fiables.

Descripción del riesgo: el código generado puede presentar fallos de seguridad, ya sean accidentales o debidos a un efecto malintencionado del modelo generador. Entre ellos puede figurar la incorporación de dependencias maliciosas; por ejemplo, los actores de amenazas pueden crear paquetes utilizando nombres inventados. El código generado puede resultar considerablemente más difícil de comprender y depurar que el código escrito a mano. La probabilidad de que surjan problemas aumenta a medida que crecen el alcance y la complejidad del código generado.

Función de seguridad: escanee el código generado por IA en busca de fallos, lógica poco segura o inclusiones maliciosas. Detecte problemas como dependencias alucinatorias o la apropiación indebida de paquetes para evitar que código comprometido o poco seguro llegue a los entornos de producción.

Seguridad de la cadena de suministro de IA

Capacidades

Uso de la IA en la sombra

El uso de herramientas de IA no aprobadas o no autorizadas por parte de empleados o departamentos, a menudo al margen de los controles oficiales de seguridad y cumplimiento normativo. Estos servicios suelen funcionar de forma externa y pueden recopilar datos empresariales confidenciales o utilizarlos para el entrenamiento de modelos sin supervisión alguna.

Descripción del riesgo: se filtran datos a terceros externos no autorizados, que pueden incluir información protegida, información financiera, información de clientes o propiedad intelectual confidencial. Es probable que los servicios no autorizados utilicen los datos enviados con fines de entrenamiento.

Función de seguridad: detecte y bloquee el uso de servicios de IA no autorizados, como herramientas de IA generativa no autorizadas o el uso de servidores MCP externos no fiables por parte de herramientas autorizadas. Un uso no supervisado podría exponer datos confidenciales y crear puntos ciegos en la supervisión de la seguridad.

Proveedores de IA que proporcionan agentes directamente

Agentes de IA externos o servicios de GenAI que se integran en los sistemas empresariales y a los que se les concede acceso interno o permisos para llevar a cabo tareas automatizadas en nombre de los usuarios o las aplicaciones.

Descripción del riesgo: es posible que, una vez iniciadas, las actividades de estos agentes no sean visibles ni controlables, ni en lo que respecta a los prompts ni en lo que se refiere a la conectividad.

Función de seguridad: supervise los agentes de IA externos a los que se les ha concedido acceso interno, ya que suelen operar con amplios privilegios y una visibilidad limitada. Aplique controles de acceso estrictos, registros y supervisión para evitar el uso indebido o acciones no deseadas dentro de los entornos internos.

La IA en los servicios a proveedores

El uso de proveedores externos o servicios de terceros que aplican la IA para procesar o analizar datos empresariales. Aquí se incluyen tanto las ofertas de IA declaradas como aquellas situaciones en las que los proveedores integran la IA en sus productos o flujos de trabajo sin comunicarlo explícitamente a los clientes.

Descripción del riesgo: es posible que los proveedores que prestan un servicio de IA conocido no protejan los datos de la empresa de forma adecuada, ya sea por motivos contractuales o debido a fallos operativos. Los proveedores pueden utilizar la IA de forma interna y discreta, eludiendo así la supervisión de la empresa que les facilita datos confidenciales.

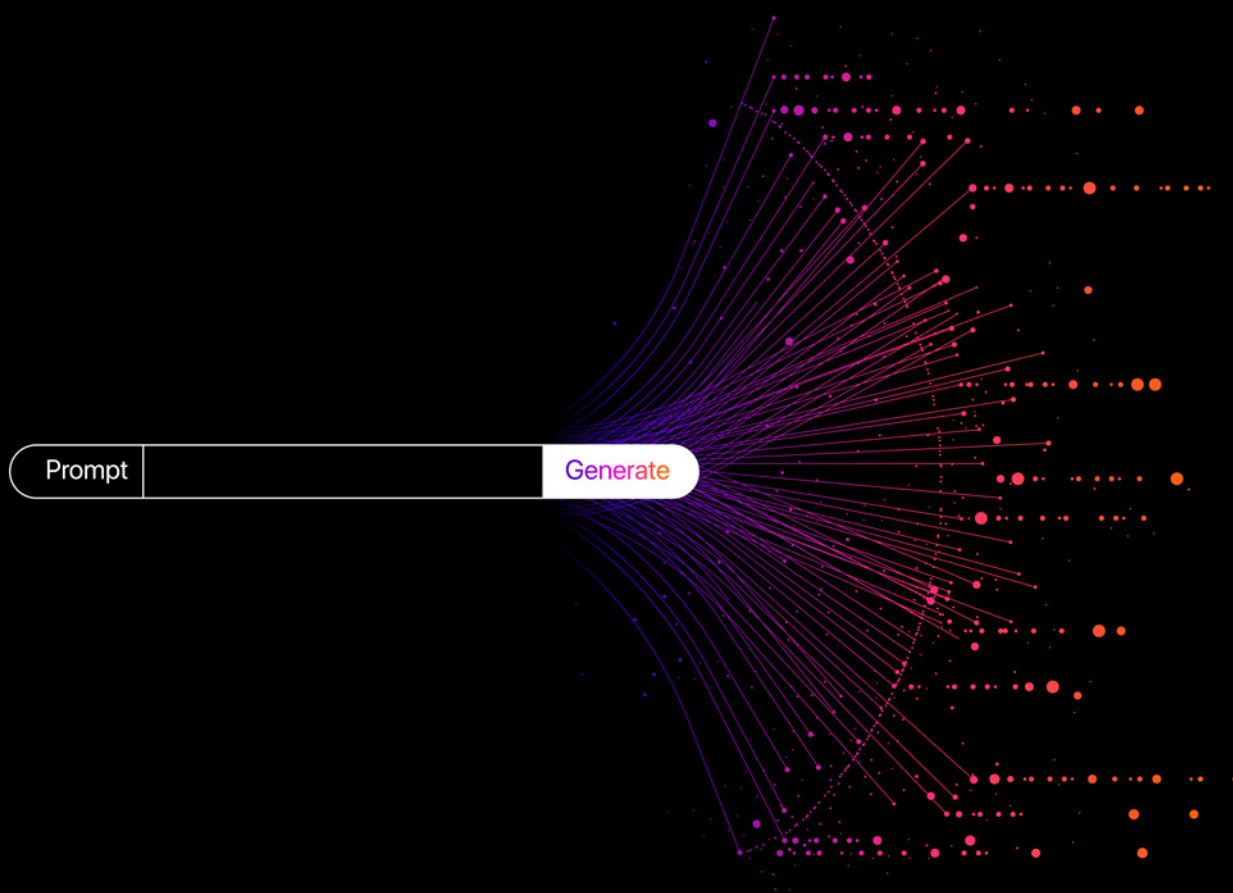
Función de seguridad: evalúe a los proveedores que utilizan la IA con los datos de su empresa (tanto los revelados como los ocultos). Verifique cómo los sistemas externos de IA procesan, almacenan y protegen dichos datos para evitar un uso no autorizado o una exposición involuntaria a través de servicios de terceros.

Escaneo de código y dependencias de agentes de IA

Proceso que consiste en examinar el código fuente de los agentes de IA y las dependencias de software asociadas a ellos con el fin de identificar vulnerabilidades, problemas de licencia o inclusiones maliciosas. De este modo se garantiza que los componentes utilizados para crear o ampliar los sistemas de IA sean seguros, cumplan con la normativa y funcionen según lo previsto.

Descripción del riesgo: Las dependencias de software pueden incluir licencias inapropiadas o haber sido comprometidas para ser maliciosas, ya sean de código abierto o comerciales. El código desarrollado para agentes de IA puede tener fallos de seguridad que reduzcan o anulen los objetivos de seguridad previstos.

Función de seguridad: lleve a cabo un análisis de seguridad estático del código interno de los agentes de IA para identificar puntos débiles y configuraciones inseguras. Revise todas las dependencias de software para detectar vulnerabilidades conocidas o inclusiones maliciosas antes de la implementación. El escaneo periódico contribuye a mantener la integridad del código y reduce el riesgo en el proceso de desarrollo de la IA.



Suministro de modelos de IA

Para empresas que desarrollan herramientas de IA

El proceso de adquisición o integración de modelos de IA externos —ya sean de código abierto, de terceros o proporcionados por un proveedor— en los sistemas o en los procesos de desarrollo de una organización.

Descripción del riesgo: los modelos procedentes de fuentes externas pueden incluir contenido e instrucciones maliciosos o incorrectos que generen resultados que lleven a los agentes de IA o a los usuarios a realizar acciones no deseadas. Podrían presentar niveles de sesgo, precisión, desviación y otras magnitudes que resulten inaceptables para la empresa o para sus casos de uso concretos. Es posible que no cuenten con las licencias adecuadas para su uso dentro de la empresa.

Función de seguridad: garantizar la seguridad de la cadena de suministro de modelos implica validar tanto su origen como su contenido. Cada modelo externo debe someterse a un análisis para detectar posibles riesgos ocultos, debe comprobarse que cuente con las licencias adecuadas y debe almacenarse de forma segura con control de versiones.

Datos de entrenamiento de IA

Para empresas que desarrollan herramientas de IA

La recopilación y preparación de conjuntos de datos utilizados para entrenar o ajustar modelos de IA o en el proceso de TEVV (pruebas, evaluación, validación y verificación). Estos conjuntos de datos pueden proceder de fuentes internas, repositorios públicos o proveedores externos, e influyen directamente en el comportamiento, la precisión y el cumplimiento normativo de los modelos.

Descripción del riesgo: los datos de entrenamiento pueden incluir datos maliciosos que provoquen que un modelo entrenado tenga efectos no deseados. Entre ellos podrían figurar datos protegidos por la legislación, datos que no cuenten con la licencia adecuada o un dataset en su conjunto que no cuente con la licencia adecuada para los usos previstos en la empresa.

Función de seguridad: la protección de los datos de entrenamiento comienza por saber exactamente de dónde proceden y qué contienen. Verifique el origen, las licencias y la integridad de los datos antes de utilizarlos. Protéjase contra la manipulación maliciosa de datos, tanto interna como externa. Identifique y gestione los riesgos relacionados con las licencias de los datos, especialmente en los modelos base, así como la posibilidad de que los datos de entrenamiento se reproduzcan directamente en los resultados del modelo.

Fortalecimiento de la preparación y la supervisión

Capacidades

Informes

El proceso de recopilar y comunicar los riesgos, los indicadores y las tendencias relacionados con la IA, con el fin de facilitar la toma de decisiones fundamentadas y la gobernanza.

Descripción del riesgo: los objetivos y las métricas son fundamentales para una gestión eficaz y la mejora constante. Sin unos informes de gestión de la IA eficaces, es posible que se pasen por alto cuestiones o tendencias críticas. El resto de la empresa podría tomar decisiones erróneas sobre la IA si no se le facilitan informes que respondan a sus intereses.

Función de seguridad: genere resúmenes claros de los incidentes de seguridad relacionados con la IA, las métricas de riesgo y el rendimiento del sistema. Haga un seguimiento de la adopción de los casos de uso previstos de la IA y señale cualquier uso no previsto o no autorizado para garantizar la supervisión y la gobernanza.

Pruebas de penetración con IA, equipos rojos y morados

Las pruebas y la validación de sistemas de IA activos mediante ataques simulados y ejercicios adversarios para identificar vulnerabilidades y lagunas en la detección.

Descripción del riesgo: los sistemas en producción suelen presentar problemas latentes que no se detectan en la configuración ni en el código. Sin pruebas de penetración, los actores de amenazas podrían descubrir vulnerabilidades específicas que los equipos de seguridad desconocen. Sin pruebas de equipos rojos o morados, podrían existir debilidades más amplias en las capacidades de exposición y detección que, sin embargo, pasarían desapercibidas.

Función de seguridad: lleve a cabo pruebas de penetración y ejercicios de «equipo rojo» o «equipo morado» para simular ataques reales contra sistemas de IA en funcionamiento. Detecte puntos débiles que puedan ser objeto de explotación, especialmente en entornos expuestos al exterior, con el fin de anticiparse a las técnicas de los adversarios y mitigarlas antes de que se produzcan.

Formación y simulacros del equipo de seguridad

Formación continua y ejercicios de simulación diseñados para desarrollar habilidades relacionadas con la detección, la investigación y la respuesta ante incidentes específicos de la IA en los equipos de seguridad.

Descripción del riesgo: una formación individual insuficiente o una falta de prácticas de respuesta ante incidentes en equipo pueden provocar fallos en la detección, investigaciones deficientes o la incapacidad de detener o eliminar un problema de ciberseguridad o un actor de amenazas.

Función de seguridad: Desarrolle competencias específicas en materia de IA y la preparación de los equipos de seguridad. La formación y las simulaciones periódicas refuerzan la capacidad para detectar incidentes relacionados con la IA, responder a ellos y recuperarse de ellos.



Un nuevo enfoque de la seguridad de la IA para los consejos de administración

La misma información que agiliza las decisiones empresariales puede, si no se protege adecuadamente, multiplicar el riesgo a una velocidad y una escala sin precedentes.

Para los responsables de seguridad, la misión no consiste únicamente en proteger los sistemas de IA, sino en traducir su complejidad en orientaciones claras y estratégicas para la dirección ejecutiva.

A nivel de junta, esto puede empezar por enmarcar la seguridad de la IA en términos de confianza, responsabilidad y resiliencia:

Confianza

Los consejos de administración desean tener la garantía de que las decisiones basadas en la IA sean precisas, interpretables, respeten la privacidad y estén a salvo de manipulaciones. Los CISO deben destacar los controles que verifican la integridad de los datos, el comportamiento de los modelos y la procedencia de cada componente de la cadena de suministro de la IA.

Responsabilidad

Aclare quién es responsable de cada nivel de riesgo relacionado con la IA: los científicos de datos, los proveedores de servicios en la nube, los proveedores y los equipos de seguridad tienen, cada uno, responsabilidades bien definidas. Los consejos de administración responden ante una estructura clara y una responsabilidad compartida, no ante la ambigüedad.

Refuerce la rendición de cuentas vinculando los principios de seguridad de la IA a marcos de gobernanza consolidados, como el Marco de Gestión de Riesgos de la IA (RMF) del NIST, la norma ISO/IEC 42001 y las directrices de la CISA y el NCSC. Esta vinculación demuestra madurez, garantiza la coherencia entre las obligaciones de cumplimiento y proporciona un punto de referencia claro para la supervisión.

Resiliencia

Haga hincapié en que la IA forma ya parte de la infraestructura crítica de la organización. Protegerla significa garantizar la continuidad ante un ataque, la auditabilidad ante un escrutinio y la agilidad ante los cambios normativos.

A medida que la IA se vaya integrando en todas las operaciones empresariales, vincular su seguridad a aspectos como la protección de la propiedad intelectual, la fiabilidad operativa y la confianza de las partes interesadas será fundamental para la resiliencia de las empresas. Las organizaciones que cuenten con sistemas de gestión que faciliten activamente la ampliación de los casos de uso de la IA, sin dejar de garantizar su seguridad, estarán en una posición óptima para innovar de forma segura, adaptarse con rapidez y mantener la confianza de los clientes, los organismos reguladores y los inversores.

Conclusión

Las mismas fuerzas que hacen que la IA sea transformadora también dificultan su seguridad cuando se integra en herramientas SaaS, infraestructura en la nube, sistemas de proveedores y aplicaciones desarrolladas dentro de la propia empresa. En este entorno, la visibilidad, la responsabilidad y la rendición de cuentas pueden difuminarse fácilmente.

Garantizar la seguridad de la IA significa recuperar esa claridad. Para ello, es necesario comprender no solo cómo se desarrolla la IA, sino también dónde opera, quién interactúa con ella y cómo influyen sus decisiones en el negocio. Al analizar los riesgos de seguridad de la IA desde esta perspectiva más amplia, que abarca las cadenas de suministro, el desarrollo, el funcionamiento en tiempo real y la gobernanza, las organizaciones pueden empezar a convertir la incertidumbre en seguridad.

El camino hacia una IA segura es un proceso continuo, pero comienza por la concienciación: saber dónde está presente la IA, cómo se comporta y cómo gestionarla de forma responsable.

Desarrollar la IA de forma segura significa hacerlo de manera responsable.

Descubra cómo armonizar la innovación con la gobernanza en nuestro documento «Hacia una IA responsable en el ámbito de la ciberseguridad»

Más información ➔



Descubra el espectro de tipos de IA en ciberseguridad

Conozca las herramientas que sustentan la resiliencia actual, desde el aprendizaje automático supervisado hasta el procesamiento del lenguaje natural (NLP), y cómo se complementan para hacer frente a las amenazas emergentes.

Más información ➔

■ **Acerca de Darktrace**

Darktrace es líder mundial en ciberseguridad basada en IA que permite a las organizaciones adelantarse cada día al panorama de amenazas en constante evolución. Fundada en 2013 en Cambridge (Reino Unido), Darktrace ofrece la plataforma de ciberseguridad esencial para proteger a las organizaciones frente a amenazas desconocidas mediante una IA que aprende de cada empresa en tiempo real. La plataforma y los servicios de Darktrace cuentan con el respaldo de más de 2.700 empleados que protegen a cerca de 10.000 clientes en todo el mundo. Para más información, visite www.darktrace.com.