La brecha de seguridad en la IA: Guía de implementación de Zero Trust para equipos de seguridad

Una guía práctica para proteger todo el ciclo de vida de la IA, desde el descubrimiento de la IA en la sombra hasta la protección del entorno de ejecución de agentes, con principios de Zero Trust, una hoja de ruta de implementación de 30/60/90 días e indicadores clave de rendimiento (KPI) para informar a la dirección.

LIBRO ELECTRÓNICO

Introducción

Los equipos de seguridad saben que tienen un problema con la IA. Lo que la mayoría aún no puede responder es cuán grande es.

¿Qué aplicaciones de inteligencia artificial (IA) se están ejecutando en su entorno en este momento? ¿Qué modelos consultan sus desarrolladores a través de los complementos del entorno de desarrollo integrado (IDE) y las herramientas de la interfaz de línea de comandos (CLI)? ¿Qué datos se transmiten a través de los prompts y alguno de sus códigos fuente, información personal identificable del cliente (PII) o contenido está sujeto a normas de divulgación regulatoria? ¿Qué agentes actúan de forma autónoma en nombre de los usuarios y a qué sistemas pueden acceder?

Esa brecha entre lo que la IA hace en tu entorno y lo que puedes ver y controlar es donde reside el riesgo real. No en escenarios de ataque teóricos, sino en el uso diario, no autorizado y bienintencionado de herramientas de IA que la seguridad nunca aprobó porque nadie se molestó en preguntar. La IA oculta y las funciones de IA integradas en plataformas de software como servicio (SaaS) ya aprobadas son las principales fuentes de esta brecha. Una herramienta que el departamento de TI aprobó para un propósito específico incluyó un asistente de IA por defecto, y ahora está procesando datos internos sin ninguna revisión de seguridad.

Los datos y las estadísticas de esta guía proceden del Informe de seguridad de IA de ThreatLabz 2026, publicado por Zscaler en 2026, y se basan en el análisis de 989,3 mil millones de transacciones de IA/ML observadas en Zscaler Zero Trust Exchange entre enero y diciembre de 2025. Todas las cifras reflejan ese período de investigación, salvo que se indique lo contrario.

Zero Trust ofrece una respuesta basada en principios:

No considere que ninguna aplicación, usuario, prompt o conector sea inherentemente fiable. Aplicar controles de acceso basados en la identidad y el contexto en todas las capas. Verificar continuamente. Minimizar el radio de impacto cuando algo sale mal. El reto reside en la secuenciación. La visibilidad debe preceder a la aplicación de las políticas. Los controles de acceso deben ser estables antes de implementar la automatización. Si se implementan de forma desordenada, se generan controles que bloquean el trabajo legítimo o que no logran detener el riesgo real.

83 %	**18.033 TB**	**39 %**	**93 %**
Crecimiento interanual de las transacciones empresariales de IA/ML: la adopción ya ha alcanzado una escala de infraestructura.	de datos transferidos a través de aplicaciones de IA en 2025, más de lo que la mayoría de los programas pueden supervisar	de transacciones de IA bloqueadas activamente: la presión de la gobernanza ya está presente	Crecimiento interanual del volumen de transferencia de datos de IA: supera la capacidad de los programas de seguridad.

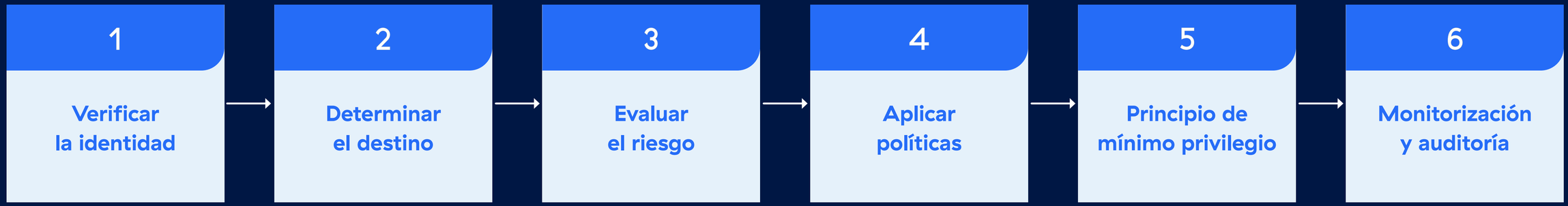
Fuente: Informe de seguridad de IA de ThreatLabz 2026, Zscaler (págs. 5—6)

Qué abarca esta guía

1. Los siete principios de Zero Trust aplicados específicamente a los sistemas, agentes y flujos de datos de IA
2. Una hoja de ruta de implementación por fases de 30/60/90 días, estructurada según el orden de dependencias
3. Cobertura de los nuevos dominios de riesgo de la IA: IA en la sombra, inyección de prompts, sistemas de agentes, tráfico del Protocolo de Contexto de Modelos (MCP) y agente-a-agente (A2A), IA integrada en plataformas SaaS y postura de seguridad de la infraestructura de IA.
4. Un modelo de KPI para informar al equipo directivo sobre el progreso del programa de seguridad de IA.

Principios de Zero Trust aplicados a la IA

Los siete principios que se detallan a continuación conforman el marco de decisión para todos los aspectos de la hoja de ruta de implementación. Cada una de ellas es una herramienta para la toma de decisiones y la evaluación de los controles. La secuencia refleja cómo un equipo de seguridad las aplicaría en la práctica.



Aplicado en secuencia. Cada principio se basa en el anterior.

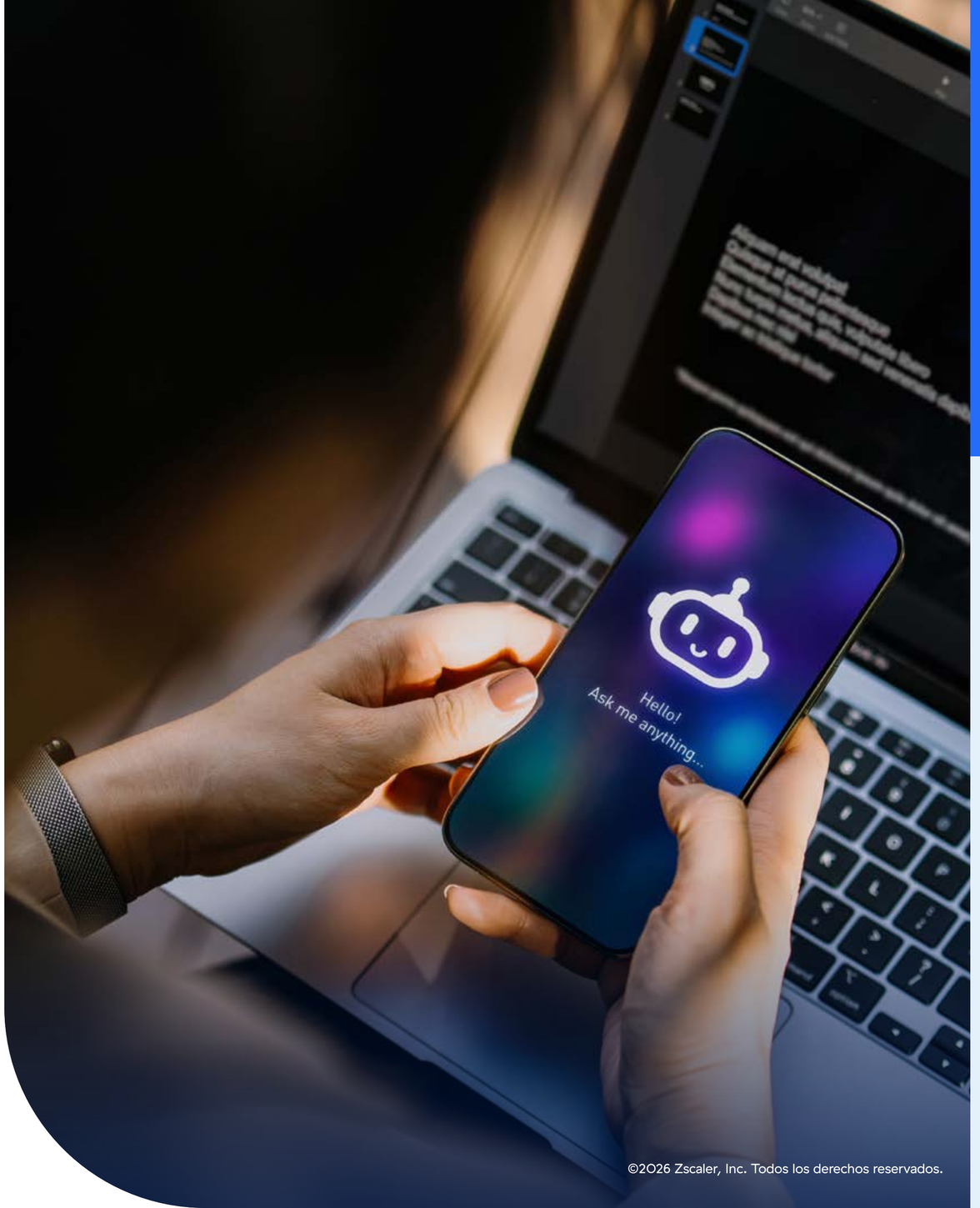
Principio 1:
VERIFICAR LA IDENTIDAD DE USUARIOS, DESARROLLADORES Y SERVICIOS.

El acceso a la IA comienza con la identidad. Todos los usuarios que accedan a una herramienta SaaS de IA generativa (GenAI), todos los desarrolladores que consulten un punto final de modelo desde un complemento de IDE y todas las cuentas de servicio que se conecten a un componente de infraestructura de IA deben autenticarse a través de un plano de control de identidad coherente. El inicio de sesión único (SSO) y la autenticación multifactor (MFA) son básicos, pero los detalles importan.

La plantilla y los desarrolladores presentan perfiles de riesgo diferentes y requieren políticas diferenciadas. Un analista de marketing que utiliza ChatGPT para la elaboración de borradores de contenido presenta riesgos de exposición de datos diferentes a los de un ingeniero de software cuyo entorno de desarrollo integrado (IDE) enruta las sugerencias de autocompletado de código a través de un complemento de GitHub Copilot conectado a un servidor MCP con acceso a repositorios internos. La autenticación de servicio a servicio para agentes de IA es el caso límite emergente para el que la mayoría de las organizaciones no están preparadas.

Controles clave

- Implementación de SSO y MFA para software como servicio (SaaS) de IA y herramientas para desarrolladores de IA.
- Políticas de identidad separadas para el uso de GenAI por parte de la fuerza laboral en comparación con las herramientas de IA para desarrolladores (IDE, CLI, integraciones de MCP).
- **Controles de cuentas de servicio para agentes de IA: principio de** mínimo privilegio, credenciales de corta duración y registros de auditoría.
- **El estado del dispositivo como señal de identidad: los dispositivos** gestionados reciben un acceso distinto al de los dispositivos personales (BYOD).



Principio 2:
DETERMINAR EL DESTINO DE LA IA AUTORIZADA Y NO AUTORIZADA

No todas las aplicaciones de IA son iguales, y un enfoque general de bloqueo total no es ni viable desde el punto de vista operativo ni eficaz en materia de seguridad. **El objetivo es una política de acceso escalonada:** las herramientas autorizadas obtienen acceso completo con inspección en línea, las toleradas obtienen acceso supervisado con restricciones de uso, y las no autorizadas o de alto riesgo se bloquean o aíslan. Esta lista de interfaces de IA es mayor de lo que la mayoría de las organizaciones prevén, multiplicándose por cuatro en un solo año a medida que se incorporan nuevas funcionalidades de IA a las plataformas existentes.

Controles clave

- Políticas de destino escalonadas para permitir, advertir, aislar o bloquear por aplicación y grupo de usuarios.
- Descubrimiento dinámico de aplicaciones para detectar nuevas herramientas de IA a medida que aparecen en el tráfico.
- Políticas independientes para GenAI SaaS, IA integrada en SaaS empresarial y herramientas para desarrolladores de IA.
- Flujo de trabajo de aprobación para solicitudes de nuevas herramientas de IA para que los usuarios tengan un camino a seguir en lugar de tener que sortear los controles.

3.400+

MÁS DE 3.400 APLICACIONES DE IA DISTINTAS DETECTADAS EN EL TRÁFICO EMPRESARIAL, CON UN CRECIMIENTO DE 4 VECES EN UN SOLO AÑO Y UN AUMENTO MENSUAL CONTINUO.

Informe de seguridad de IA de ThreatLabz 2026, Zscaler (pág. 5)

Principio 3:
EVALUAR EL RIESGO CON DECISIONES BASADAS EN EL CONTEXTO.

Las decisiones de acceso a la IA nunca deberían ser binarias. El acceso de un mismo usuario a la misma herramienta de IA presenta riesgos diferentes según el contexto: dispositivo, sensibilidad de los datos y acción realizada. Una evaluación eficaz del riesgo requiere definir de antemano las categorías de datos sensibles: código fuente, información de identificación personal (PII), datos del sector de las tarjetas de pago (PCI), información médica protegida (PHI), resultados financieros no publicados y detalles confidenciales sobre fusiones y adquisiciones (M&A). Las principales categorías de infracciones detectadas en el tráfico empresarial indican con exactitud qué información envían los empleados: nombre y datos de identificación nacional, código fuente e historiales médicos.

Controles clave

- Categorías explícitas de datos sensibles para el código fuente, PII/PCI/PHI, secretos y planes no publicados
- Puntuación de riesgo por grupo de usuarios, departamento, estado del dispositivo y sensibilidad de los datos.
- Ajuste dinámico de políticas basado en señales contextuales, no solo en reglas estáticas de permitir o bloquear
- Tratamiento de riesgos diferenciado para las interacciones con IA que incluyen cargas de archivos frente a las que solo utilizan mensajes de texto.

MÁS DE 410 MILLONES

INFRACCIONES DE PREVENCIÓN DE PÉRDIDA DE DATOS (DLP) VINCULADAS ÚNICAMENTE A CHATGPT. TIPOS PRINCIPALES: FUGA DE NOMBRES, NÚMEROS DE SEGURIDAD SOCIAL, CÓDIGO FUENTE, DATOS MÉDICOS.

Informe de seguridad de IA de ThreatLabz 2026, Zscaler (págs. 6, 15).

Principio 4:
APLICAR LA POLÍTICA EN LÍNEA, POR SESIÓN.

Los registros muestran lo que ha ocurrido. Para cuando encuentre un registro que demuestre que un desarrollador envió código fuente propietario a un modelo no autorizado, los datos ya se habrán perdido. La mayoría de las organizaciones agravan este problema al implementar DLP para la carga de archivos en herramientas de IA, mientras dejan sin inspeccionar el texto de los prompts. Un desarrollador que introduce código fuente directamente en una ventana de prompt elude esos controles sin activar una sola alerta. El crecimiento de la transferencia de datos demuestra la magnitud de lo que se mueve a través de esos canales desprotegidos.

Controles clave

- DLP en línea para prompts, respuestas y cargas de archivos, no solo a nivel de URL o categoría.
- Acciones de permitir, advertir, censurar, bloquear y aislar disponibles en los puntos de aplicación de políticas.
- Clasificación de prompts para las principales aplicaciones de GenAI con detección contextual.
- Moderación de contenido para interacciones fuera de tema, restringidas y con información sensible desde el punto de vista competitivo.
- Aislamiento del navegador para herramientas de IA permitidas, pero de riesgo: acceso sin confianza implícita.

93 %

CRECIMIENTO INTERANUAL DEL VOLUMEN DE TRANSFERENCIA DE DATOS DE IA EMPRESARIAL, SUPERIOR A LA CAPACIDAD DE SEGUIMIENTO Y GOBERNANZA DE LA MAYORÍA DE LOS PROGRAMAS DE SEGURIDAD.

Informe de seguridad de IA de ThreatLabz 2026, Zscaler (págs. 5 y 14).

Principio 5:
MINIMIZAR EL ACCESO Y LA EXPOSICIÓN

Los sistemas de IA crean vías de acceso que a menudo quedan fuera de los modelos tradicionales de identidad y red. Un proceso de generación aumentada por recuperación (RAG, por sus siglas en inglés) puede proporcionar a un modelo acceso a un corpus de documentos. Un agente puede conectarse a herramientas internas, conjuntos de datos, almacenes de vectores o servidores MCP. Esos permisos son tan reales como el acceso de cualquier usuario, aunque no aparezcan en las revisiones de acceso estándar.

Minimizar el acceso y la exposición implica reducir el alcance de los sistemas de IA y limitar la forma en que se comunican sus componentes. Los agentes, las canalizaciones de datos, los componentes RAG y las fuentes de datos de backend solo deben conectarse a través de rutas aprobadas explícitamente. Sin un enfoque y una segmentación basados en el principio de mínimo privilegio, un agente con permisos excesivos o un componente de IA comprometido puede convertirse en una vía de movimiento lateral.

Controles clave

- Definición del alcance del acceso basado en roles para aplicaciones de IA, agentes, conectores de modelos, conjuntos de datos, almacenes de vectores y componentes de infraestructura de IA.
- Revisión periódica y reducción del acceso excesivo a conjuntos de datos sensibles y sistemas conectados a IA.
- Conjuntos mínimos de permisos para agentes de IA: definir a qué pueden acceder y restringir todo lo demás.
- Segmentación de servicios de IA, plataformas de agentes, flujos de datos, componentes RAG y fuentes de datos de backend
- Restringir el acceso este–oeste entre los componentes de IA a rutas definidas explícitamente.
- Inspección y registro con reconocimiento de protocolo para el tráfico MCP y A2A

Minimizar el radio de impacto en cada ruta de acceso de la IA.

Considere cada conector de IA, permiso de agente, canalización RAG, almacén de vectores, servidor MCP y conexión entre componentes como un punto de exposición. Defina el alcance de sus accesos, segmente las zonas de comunicación, inspeccione el tráfico y registre la actividad.

Principio 6:
MONITORIZACIÓN CONTINUA Y CAPACIDAD DE AUDITORÍA.

La verificación continua es el modelo operativo, no un hito que se alcanza y después se deja atrás. En el caso de los sistemas de IA, esto implica el registro de prompts y respuestas, alertas en tiempo real sobre infracciones de políticas, paneles de uso y un flujo de trabajo de investigación que no requiera correlacionar manualmente registros procedentes de cinco herramientas distintas. **La presión regulatoria que impulsa este enfoque es real:** la Ley de IA de la Unión Europea (EU AI Act), el Marco de Gestión de Riesgos de IA del Instituto Nacional de Estándares y Tecnología (NIST AI RMF) y los marcos emergentes de gobernanza de la IA exigen controles demostrables y documentados.

Controles clave

- Registro de prompts y respuestas para las interacciones de IA en línea.
- Paneles de uso por usuario, departamento y aplicación de IA con superposiciones de señales de riesgo.
- Flujo de trabajo de alerta e investigación para infracciones de políticas y anomalías relacionadas con la IA.
- Informes listos para auditoría: qué IA se utiliza, qué controles se aplican y qué infracciones se produjeron.
- Alineación de la gobernanza con NIST AI RMF, EU AI Act y otros marcos aplicables.

1	2	3	4	5
Reactivo	**Visible**	**Gobernado**	**Automatizado**	**Continuo**
Sin inventario Bloqueo ad hoc	Huella de IA conocida Políticas de referencia	Políticas basadas en roles Flujos de trabajo de gobernanza	Orientación automatizada Cumplimiento continuo	Red teaming y Protecciones en tiempo de ejecución

Madurez del programa →

Días 1—30

Establecer la visibilidad

No se puede proteger una IA que no se puede ver. Los primeros 30 días se centran en establecer un inventario suficiente y unos controles básicos que permitan tomar decisiones fundamentadas, no en bloquearlo todo.

DÍAS 1—30 Establecer la visibilidad	DÍAS 31 — 60 Desarrollar la madurez de las políticas	DÍAS 61 — 90 Automatizar y validar
• Implementar el descubrimiento de aplicaciones de IA • Definir las categorías de datos sensibles • Políticas básicas de acceso • Primer panel para la dirección	• Políticas de acceso basadas en roles • Poner en marcha los flujos de trabajo de gobernanza • Implementación de AI-SPM • Protecciones para la IA privada	• Formación automatizada para los usuarios • Cumplimiento continuo • Programa periódico de ejercicios de red teaming • Protecciones de circuito cerrado

Hoja de ruta de implementación de 30/60/90 días de un vistazo

Lista de verificación de implementación de 30 días

- **Establezca una cadencia periódica de revisión de las herramientas de IA:** nuevas herramientas detectadas en el tráfico, solicitudes de excepción recibidas y cambios introducidos en las políticas
- Cree el primer panel de seguridad de IA para la dirección: huella de las aplicaciones de IA (autorizadas frente a IA en la sombra), eventos de DLP por categoría, acciones de las políticas y tendencias
- Defina procedimientos de respuesta para los incidentes de seguridad relacionados con la IA más habituales: datos sensibles en los prompts, acceso a destinos bloqueados y cargas de archivos de gran tamaño en herramientas de IA
- **Centralice los registros de seguridad de IA:** actividad de los usuarios, uso de las aplicaciones, eventos de prompts y respuestas, acciones de las políticas e infracciones, todo ello en un único panel

Operativizar

- **Revise y ajuste las políticas de DLP utilizando las dos primeras semanas de telemetría:** reduzca los falsos positivos y mejore la detección de los patrones confirmados de exposición de datos
- **Habilite la moderación de contenido para aplicar la política de uso aceptable:** interacciones fuera de tema, temas restringidos, contenido sensible desde el punto de vista competitivo y respuestas inapropiadas
- **Añada aislamiento del navegador para las herramientas de IA del nivel "permitidas, pero de riesgo": los** usuarios pueden acceder a la herramienta, pero las cargas y descargas de archivos, así como las operaciones de copiar y pegar, quedan restringidas en la capa del navegador
- **Amplíe la cobertura de DLP para abarcar todas las categorías de datos sensibles:** código fuente (más lenguajes y patrones), PII (más allá de los formatos evidentes), datos de titulares de tarjetas de pago (PCI), PHI y contenido empresarial confidencial

Ampliar la protección de datos

- **Publique una guía de uso aceptable de la IA basada en un semáforo:** verde (autorizado), amarillo (uso con restricciones) y rojo (bloqueado, con su justificación). Haga visible el proceso de solicitud de excepciones
- Habilite DLP en línea para los prompts, comenzando por reglas de detección de alta fiabilidad para las categorías de datos sensibles: patrones de código fuente, formatos de credenciales y esquemas de PII
- **Restrinja las acciones de mayor riesgo en las herramientas no autorizadas:** cargas de archivos, operaciones de copiar y pegar de gran volumen y descarga de resultados de IA a almacenamiento no gestionado
- **Aplique la política de destino:** advierta o bloquee, por grupo de usuarios y aplicación, el acceso a destinos de IA claramente no autorizados o de alto riesgo.

Aplicar controles básicos

- **Establezca un inventario básico de riesgos de IA:** qué aplicaciones acceden a qué tipos de datos y qué grupos de usuarios hacen un mayor uso de la IA.
- **Defina explícitamente las categorías de datos sensibles:** código fuente, PII, datos PCI, PHI, secretos y credenciales, datos financieros no publicados y planes estratégicos o de fusiones y adquisiciones (M&A) confidenciales

Defina qué debe protegerse

- Habilite la visibilidad de los prompts en las aplicaciones de IA con mayor volumen de tráfico para comprender qué datos circulan a través de los prompts y las respuestas
- Determine qué herramientas de IA tienen integraciones con sistemas internos y a qué tipos de datos acceden
- **Detecte la IA en la sombra:** herramientas en uso activo que nunca fueron aprobadas formalmente
- Implemente el descubrimiento de aplicaciones de IA para inventariar las aplicaciones GenAI SaaS, la IA integrada en plataformas empresariales y las herramientas de desarrollo de IA utilizadas en todo el entorno

ASPECTO QUE DEBE TENERSE EN CUENTA

Comenzar con políticas de advertencia, en lugar de bloqueos estrictos, reduce la fricción y permite detectar falsos positivos en las reglas de DLP antes de que empiecen a bloquear el trabajo legítimo. El inconveniente es que durante la fase de advertencia pueden seguir produciéndose incidentes de exposición de datos. Equilibre este enfoque con su tolerancia al riesgo respecto a los datos sensibles. Si la fuga de código fuente constituye un riesgo crítico, merece la pena asumir inicialmente los falsos positivos derivados de aplicar bloqueos estrictos basados en la detección de patrones de código.

Entregables en 30 días

- Inventario de aplicaciones de IA con clasificación por niveles de riesgo (autorizadas, toleradas, prohibidas)
- Las políticas iniciales de acceso están implantadas para las aplicaciones de IA y los grupos de usuarios de mayor riesgo.
- La DLP para prompts está activa para las categorías de datos sensibles y su ajuste inicial se ha completado.
- Primer panel para la dirección con la huella de IA y las métricas de riesgo.
- Se han publicado la guía de uso aceptable de la IA y el proceso de solicitud de excepciones.

Días 31—60

Desarrollar la madurez de las políticas

Los días 31 a 60 se centran en la visibilidad y los controles iniciales para establecer un modelo de políticas maduro, flujos de trabajo de gobernanza funcionales y un inventario de infraestructura de IA.

Lista de verificación de implementación de 60 días

Desarrolle el modelo de políticas de acceso.

- **Establezca políticas de acceso a la IA específicas para cada departamento:** Recursos Humanos, Finanzas, Ingeniería y Ventas tienen distintas necesidades en cuanto a herramientas de IA y diferentes perfiles de riesgo de exposición de datos
- **Implemente controles condicionales por capas:** niveles de acceso para dispositivos gestionados frente a BYOD, señales de acceso geográfico y evaluación del riesgo de la postura del dispositivo
- **Formalice tres niveles de acceso para las aplicaciones de IA:** autorizado (acceso completo con inspección en línea), tolerado (acceso supervisado con restricciones de uso) y prohibido (bloqueado con procedimiento de excepción)
- Revise y retire las excepciones concedidas durante los primeros 30 días conforme al modelo de políticas, ahora más maduro

Establecer procesos de gobernanza

- **Asigne los requisitos de gobernanza de la IA a las políticas internas y a los marcos aplicables:** NIST AI RMF, EU AI Act e HIPAA, cuando la PHI esté dentro del alcance
- **Formalice el flujo de trabajo de aprobación de herramientas de IA:** presentación, revisión de seguridad, clasificación del riesgo, aprobación condicional y revalidación periódica
- Establecer etiquetas de sensibilidad estándar para los tipos de datos relevantes para el uso de la IA, con reglas de manejo claras para cada etiqueta
- **Defina las vías de escalado para los incidentes de seguridad relacionados con la IA:** quién es responsable de la respuesta, qué constituye un incidente que requiere notificación a la dirección y qué evidencias deben conservarse

Inventaríe la infraestructura de IA.

- Inventariar modelos de IA, agentes, servicios, conjuntos de datos, almacenes de vectores y canalizaciones, y construir la lista de materiales de IA (AI-BOM) con linaje desde los conjuntos de datos hasta su uso en tiempo de ejecución
- **Ejecute una evaluación de AI-SPM:** identifique configuraciones incorrectas, permisos excesivos, exposición de datos sensibles y vulnerabilidades en los componentes de la infraestructura de IA
- **Priorice la remediación de los riesgos más graves detectados por AI-SPM:** permisos excesivos, exposición de datos sensibles y rutas de acceso sin restricciones a los componentes de IA
- **Valide la cobertura en las principales plataformas de IA:** modelos alojados en la nube, plataformas de agentes implementadas en la empresa, integraciones con servidores MCP e infraestructura de IA no gestionada

Proteja las aplicaciones de IA privadas

- **Añada detección de prompt injection a los controles de las aplicaciones de IA privadas:** entradas diseñadas por atacantes para anular las instrucciones del modelo o extraer información confidencial
- **Implemente la supervisión del envenenamiento de datos en las canalizaciones RAG:** detecte modificaciones no autorizadas del corpus documental que puedan sesgar las respuestas del modelo
- **Habilite el registro de prompts y respuestas para los sistemas de IA privados con detección de anomalías:** patrones de consulta inusuales, contenido de salida inesperado e interacciones fuera de ámbito
- Establezca flujos de trabajo para revisar las interacciones del modelo en las aplicaciones de IA privadas de alta sensibilidad

Entregables a los 60 días

- Modelo maduro de políticas de acceso basadas en roles con aplicación condicional en funcionamiento.
- Flujo de trabajo operativo de gobernanza de la IA: aprobación de herramientas, etiquetado de sensibilidad y vías de escalado definidas.
- Línea base de visibilidad de AI-SPM completada con un plan priorizado de remediación.
- AI-BOM que cubre los principales componentes de la infraestructura de IA.
- Protecciones de IA privada implantadas para las aplicaciones internas de IA de mayor prioridad
- Periodicidad de los informes de cumplimiento definida y primer informe elaborado.

Días 61—90

Automatizar y validar

Durante los días 61—90, el programa pasa a ser autosostenible: aplicación automatizada de políticas, supervisión continua del cumplimiento y validación mediante red teaming.

Lista de comprobación de la implementación en 90 días

Automatice la aplicación de las políticas.

- **Implemente orientación automatizada para los usuarios en caso de infracciones de las políticas:** advierta con orientación específica para el contexto antes de escalar al bloqueo
- **Integre la gestión de incidencias mediante ITSM o SOAR para las infracciones reiteradas:** cree automáticamente incidencias de investigación para los usuarios que activen repetidamente los mismos controles
- **Configure desencadenadores automáticos de cuarentena y aislamiento para comportamientos de alto riesgo: envío de** grandes volúmenes de datos confidenciales, acceso a destinos de IA prohibidos y patrones de prompts anómalos
- **Cree mecanismos de retroalimentación:** supervise las tendencias de incumplimiento de las políticas a lo largo del tiempo para identificar comportamientos que indiquen la necesidad de ajustar las políticas, en lugar de reforzar su aplicación

Consolide la elaboración de informes de cumplimiento.

- **Configure la supervisión continua de la postura de cumplimiento de la IA:** qué controles están activos, qué activos están cubiertos y dónde existen lagunas de cobertura
- **Asigne los controles a los marcos aplicables:** NIST AI RMF (incluido NIST AI 600–1 para los riesgos de la IA generativa), EU AI Act, GDPR e HIPAA cuando la PHI esté dentro del alcance
- Cree paquetes de informes preparados para auditoría: resumen del uso de las aplicaciones de IA, acciones y resultados de la DLP para prompts, excepciones de política concedidas y expiradas, y estado de la remediación de los hallazgos de AI–SPM
- **Establezca una periodicidad trimestral para revisar el programa de seguridad de la IA: rendimiento de las** políticas, lagunas de cobertura, nuevas áreas de riesgo y ajustes de la hoja de ruta

Reforzar y validar mediante red teaming

- **Establezca una cadencia de red teaming para las aplicaciones de IA desarrolladas internamente:** bibliotecas predefinidas de pruebas de ataque y casos de prueba personalizados orientados al contexto específico de su aplicación
- **Ejecute pruebas que abarquen toda la superficie de ataque de la IA:** prompt injection, jailbreaks, envenenamiento del contexto, extracción de PII, fuga de código fuente, comportamiento fuera de tema y regresión de referencia del modelo
- **Implemente políticas de protección en tiempo de ejecución para los sistemas de IA en producción:** detectores de prompt injection, fugas de datos (PII y código fuente), intentos de jailbreak e infracciones de las políticas de moderación de contenido
- **Automatice la conversión de los hallazgos del red teaming en barreras de seguridad en tiempo de ejecución:** la brecha entre las pruebas y la aplicación en producción es donde fracasan la mayoría de los programas de seguridad para aplicaciones de IA

16 min.	El 72 %	El 100 %
Tiempo medio hasta el primer fallo crítico en las pruebas de red teaming de IA.Más de 222.000 ataques adversarios ejecutados en más de 25 entornos.	de las organizaciones detectaron una vulnerabilidad crítica ya en la primera prueba.El riesgo crítico está presente desde el primer día, incluso en entornos maduros.	de los sistemas de IA evaluados presentaban al menos una vulnerabilidad crítica. Ningún sistema de IA es seguro por defecto. Las pruebas continuas son imprescindibles.

Validar la alineación de la arquitectura

- **Aplique los seis principios de Zero Trust a todas las rutas de acceso a la IA:** verifique que cada principio se aplique en todas las superficies de acceso de los usuarios, los desarrolladores y la infraestructura de IA.
- **Identifique y subsane las lagunas de las políticas en los puntos de intersección:** lugares donde termina una capa de aplicación y comienza otra sin una cobertura coherente.
- Estandarice la aplicación de las políticas mediante un modelo unificado que abarque la verificación de identidad, el control de destinos, la evaluación de riesgos, la aplicación en línea, la minimización del acceso y la exposición, y la supervisión.

Entregables a los 90 días

- Flujos de trabajo automatizados de respuesta a infracciones y orientación a los usuarios en funcionamiento.
- Supervisión continua del cumplimiento operativa, con la alineación con los marcos documentada.
- Paquete de informes de gobernanza para la dirección y las auditorías.
- Cadencia de red teaming para la IA establecida, con políticas de protección en tiempo de ejecución activas.
- Conversión de los hallazgos del red teaming en barreras de seguridad de producción mediante un circuito cerrado.
- Cuadro de mando del programa y hoja de ruta trimestral para la gobernanza continua.

Cómo Zscaler permite la adopción segura de la IA

La plataforma Zero Trust Exchange™ se creó para gestionar y proteger la IA a escala empresarial. Amplía la misma plataforma que ya gobierna el acceso de usuarios, cargas de trabajo y redes para miles de empresas de todo el mundo, y no es una adaptación de una arquitectura de seguridad heredada.

Cada etapa del ciclo virtuoso se corresponde con un conjunto específico de capacidades de la plataforma. La tabla que aparece a continuación muestra lo que Zscaler ofrece en cada capa.

DESCUBRA Conozca toda su huella de IA	**CONTROL** Aplique el acceso Zero Trust a la IA	**PROTEGER** Reforzar y proteger las aplicaciones de IA
Gestión de activos de IA Zscaler descubre todas las aplicaciones, modelos, agentes, servidores MCP y funciones de IA integradas en aplicaciones SaaS de su entorno, incluidas las herramientas que el departamento de TI nunca aprobó.	**Seguridad de acceso mediante IA** Aplique controles de acceso Zero Trust para la IA SaaS, la IA integrada en plataformas empresariales y los entornos de desarrollo de IA, con inspección en línea en todas las capas.	**Red Teaming para IA y barreras de seguridad para IA** Ponga a prueba las aplicaciones de IA desarrolladas internamente frente a condiciones adversarias reales y convierta automáticamente los hallazgos en políticas de protección en tiempo de ejecución.
• Descubrimiento de IA en la sombra y visibilidad de su uso en toda la organización. • **Lista de materiales de IA (AI-BOM):** linaje desde los conjuntos de datos hasta los modelos, los agentes y el uso en tiempo de ejecución. • **AI-SPM:** evaluación continua de la postura para detectar configuraciones incorrectas, permisos excesivos y exposición de datos sensibles. • Detección de agentes de IA implementados en la empresa e integrados en aplicaciones SaaS.	• Políticas granulares de permitir, advertir, aislar y bloquear por grupo de usuarios, aplicación y tipo de datos. • DLP en línea para prompts, respuestas y cargas de archivos en todo el tráfico de IA. • Zero Trust Browser para el acceso desde dispositivos no gestionados y BYOD a herramientas de IA. • Extracción y clasificación de prompts para las principales aplicaciones GenAI.	• **Red teaming automatizado y continuo:** prompt injection, jailbreaks, envenenamiento del contexto, fuga de datos y desviación del comportamiento. • Refuerzo de prompts y evaluación comparativa de modelos, con correspondencia de gobernanza con NIST AI RMF, EU AI Act y OWASP LLM Top 10. • **Barreras de seguridad en tiempo de ejecución:** detectores de jailbreaks, fugas de PII y código fuente, moderación de contenido y comportamientos fuera de tema. • **Generación de políticas en circuito cerrado:** los hallazgos del red teaming se convierten automáticamente en detectores para producción.

LA VENTAJA DE LA PLATAFORMA

La mayoría de los proveedores de seguridad resuelven solo una parte del problema de la seguridad de la IA. Zscaler abarca todo el ciclo de vida de la IA en una única plataforma: descubre la IA en todo el entorno, controla quién puede acceder a cada herramienta de IA mediante políticas Zero Trust y protege las aplicaciones y la infraestructura de IA desde el desarrollo hasta la ejecución. El circuito cerrado entre los hallazgos del red teaming y las barreras de seguridad en tiempo de ejecución en producción es una capacidad que ninguna solución puntual puede replicar.

Preguntas frecuentes

¿Qué significa en la práctica la IA Zero Trust?
La IA Zero Trust consiste en aplicar el principio fundamental de Zero Trust —no asumir ninguna confianza implícita y verificar continuamente— a todas las interacciones con la IA en su entorno. En la práctica, ninguna aplicación de IA, usuario, dispositivo, prompt o conector es intrínsecamente fiable. El acceso se concede en función de la identidad verificada, la postura del dispositivo, la sensibilidad de los datos y el contexto de la sesión. Los controles se aplican en línea y en tiempo real. Todas las decisiones de acceso quedan registradas y pueden auditarse.

Empiece por el descubrimiento. No se puede gobernar la IA que no se ve, y ThreatLabz rastrea más de 3.400 aplicaciones de IA en el tráfico empresarial, una cifra que se ha cuadruplicado en un solo año, impulsada en gran medida por las funciones de IA integradas en herramientas que el departamento de TI ya había aprobado. Una vez que disponga de visibilidad, clasifique las aplicaciones de mayor riesgo y establezca políticas de destino de referencia. Añada DLP para prompts en las categorías de datos de mayor riesgo. Publique una guía transparente de uso aceptable y un proceso de solicitud de excepciones para que los usuarios dispongan de una vía conforme a las políticas, en lugar de verse incentivados a eludir los controles.

¿Cómo protegen los controles de acceso a la IA y la DLP en línea los prompts y las cargas de archivos? Los controles de acceso determinan quién puede acceder a qué herramientas de IA, desde qué dispositivos y ubicaciones, y en qué condiciones. La DLP en línea actúa en la capa de contenido e inspecciona en tiempo real lo que los usuarios envían realmente: el texto de los prompts, las cargas de archivos y las interacciones de copiar y pegar. Ambos controles son complementarios: el control de acceso limita quién puede acceder a cada herramienta de IA, mientras que la DLP controla qué datos pueden circular por esos canales autorizados. Una brecha en cualquiera de ellos genera exposición, por lo que las organizaciones que implementan DLP para las cargas de archivos, pero omiten la inspección de los prompts, siguen asumiendo un riesgo significativo de fuga de datos.

¿Qué es AI-SPM y por qué es importante para la adopción de la IA?
AI-SPM es la evaluación continua de las configuraciones, los permisos y los riesgos de exposición de la infraestructura de IA; el equivalente en IA de CSPM para la infraestructura en la nube. Descubre los activos de IA, evalúa sus configuraciones conforme a las mejores prácticas de seguridad, identifica permisos excesivos y exposición de datos sensibles, y supervisa la evolución de la postura de seguridad a lo largo del tiempo. Es importante porque la superficie de configuraciones incorrectas en la infraestructura de IA es amplia y poco comprendida por la mayoría de las organizaciones que pasan de la experimentación al despliegue de la IA.

¿Qué KPI demuestran mejor a la dirección que el riesgo asociado a la IA está disminuyendo sin afectar a la productividad?
El enfoque más eficaz combina una métrica de reducción del riesgo con una métrica de adopción, para que la dirección pueda observar ambas tendencias de forma conjunta. La disminución del número de aplicaciones de IA en la sombra, junto con el aumento de la adopción de herramientas de IA autorizadas, es una clara señal de que la gobernanza está funcionando. Las acciones de la DLP para prompts (distribución entre bloquear, advertir y permitir) muestran cómo se está calibrando la aplicación de las políticas. El tiempo de respuesta a las solicitudes de excepción y el tiempo de aprobación de herramientas de IA reflejan la capacidad de respuesta del proceso de gobernanza.

Actúe con rapidez. Manténgase seguro.

Acerca de Zscaler

Zscaler (NASDAQ: ZS) es pionera y líder mundial en seguridad Zero Trust. Las mayores empresas del mundo, organizaciones de infraestructuras críticas y organismos gubernamentales confían en Zscaler para proteger usuarios, sucursales, aplicaciones, datos y dispositivos, y para acelerar sus iniciativas de transformación digital. Distribuida en más de 160 centros de datos en todo el mundo, la plataforma Zscaler Zero Trust Exchange™, combinada con IA avanzada, combate cada día miles de millones de ciberamenazas e infracciones de políticas y ayuda a las empresas modernas a aumentar la productividad reduciendo los costes y la complejidad.

+1 408.533.0288 **Zscaler, Inc. (HQ) • 120 Holger Way • San Jose, CA 95134** **zscaler.com/es**