FortressAI y la seguridad de la IA generativa

Cómo FortressAI gobierna el uso de la IA generativa a nivel de kernel, deteniendo las fugas de datos hacia ChatGPT, Copilot, Gemini y otras herramientas de IA antes de que ocurran.

¿Qué es FortressAI?

Respuesta breve: FortressAI es el producto de protección de datos de IA generativa de Cyber Crucible. Aplica la política en la capa más profunda del sistema operativo, verificando el acceso a los datos en el dispositivo y bloqueando, redactando o permitiendo de forma segura la información antes de que pueda llegar a ChatGPT, Copilot, Gemini o cualquier otra herramienta de IA.

El problema que resuelve

Las herramientas públicas de IA están transformando la forma en que las personas trabajan, y al mismo tiempo abren una nueva vía para que los datos salgan de una organización. Cada indicación (prompt) conlleva un riesgo. Los empleados con frecuencia comparten información sensible sin darse cuenta de que han hecho algo incorrecto, y las herramientas de seguridad tradicionales nunca fueron diseñadas para esto.

Por qué la aplicación a nivel de kernel es importante aquí

FortressAI opera a la misma profundidad que el resto de la plataforma Cyber Crucible. Esto significa que protege contra *cualquier* herramienta de IA — incluidas aquellas que aún no existen — en lugar de mantener una lista de aplicaciones específicas para vigilar. No necesita entender un nuevo producto de IA para impedir que los datos sensibles lleguen a él.

Debido a que la aplicación es local, el contenido de las indicaciones nunca se envía a un servicio en la nube para su inspección. La protección no crea la exposición que pretende prevenir.

FortressAI no es un motor separado añadido posteriormente. Es una capa de política sobre el mismo análisis conductual de la memoria del kernel que impulsa la protección de datos e identidad, lo cual es lo que le permite evaluar no solo si una herramienta de IA está aprobada, sino también si ha sido manipulada.

¿Qué es la Shadow AI y quién dentro de mi organización está generando el riesgo?

Respuesta breve: Shadow AI se refiere a empleados que utilizan herramientas de IA que TI no ha aprobado o de las que no tiene conocimiento. El riesgo no se limita al personal de base: los administradores con privilegios y las cuentas comprometidas suelen representar la mayor exposición.

Tres fuentes de exposición

- **Empleados** que utilizan herramientas de IA no autorizadas para la productividad cotidiana.
- **Administradores con privilegios**, cuyo acceso de alto nivel implica que cualquier información que introduzcan en una herramienta de IA puede incluir datos corporativos altamente sensibles.
- **Cuentas comprometidas**, en las que un atacante con credenciales válidas utiliza herramientas de IA como canal de exfiltración, generando un tráfico que parece productividad habitual.

Esto ya está ocurriendo

Cyber Crucible ha observado directamente accesos masivos a grandes volúmenes de datos a la vez por parte de empleados que utilizan herramientas de IA para buscar y cargar información. No se trata de un riesgo futuro teórico que se esté modelando; es un comportamiento visible en implementaciones reales actualmente.

La tercera categoría merece especial atención. Un atacante que utiliza herramientas de IA para extraer datos es difícil de distinguir de un empleado entusiasta, a menos que la aplicación de controles se realice en la capa de acceso a los datos, donde la intención y la política pueden evaluarse independientemente de quién haya iniciado sesión.

¿Por qué DLP, los firewalls de IA y los LLM privados no logran detener las fugas de datos con IA?

Respuesta breve: Cada uno aborda parte del problema desde la capa equivocada. El DLP puede eludirse, los firewalls de IA se sitúan en el borde de la red y son ciegos a la actividad en el dispositivo, y los LLM privados son costosos sin resolver nada respecto a los empleados que igualmente usan herramientas públicas.

Dónde falla cada enfoque

- Los **LLM privados** son costosos de crear y mantener, y no impiden que nadie abra ChatGPT en una pestaña del navegador de todos modos.
- Los **firewalls de IA** inspeccionan el tráfico de red, lo que los deja ciegos ante lo que ocurre en el propio endpoint y ante la actividad interna.
- El **DLP tradicional** fue diseñado para archivos y flujos de correo electrónico, y es eludido rutinariamente por vías que nunca fue concebido para vigilar.

El problema de la lógica invertida

Varios enfoques de esta categoría intentan proteger los datos frente a la IA insertando *otra* IA para filtrar la salida. Esto añade costo y una nueva dependencia sin abordar el punto real donde los datos salen: el endpoint, en el momento del acceso.

FortressAI aplica su protección a nivel del kernel, por debajo de todas estas capas, y se aplica sin importar qué herramienta, navegador o extensión esté involucrada.

¿Cómo protege FortressAI los datos en un navegador web?

> **Capacidad en beta.** La detección y aplicación de políticas de uso de IA basada en navegador se encuentra actualmente en fase beta. Es funcional y está en uso activo, pero evalúela en su entorno antes de depender de ella como control.

Respuesta breve: FortressAI supervisa la actividad del navegador desde el kernel y la evalúa conforme a la política. Si el navegador, un sitio web o una extensión intenta acceder a datos fuera de la política, el permiso se revoca en la capa del kernel; y si el navegador muestra señales de explotación, se revocan todos los derechos de acceso a los datos y se suspende el proceso.

La secuencia

1. Un usuario abre un navegador web — Edge, Chrome o Firefox.
2. La supervisión a nivel de kernel evalúa el navegador y cualquier actividad de página o extensión conforme a la política de FortressAI asignada.
3. Si el navegador, el sitio o la extensión intenta acceder a datos fuera de la política, el permiso se revoca en la capa del kernel.
4. Si el navegador muestra señales de explotación, se revocan todos los derechos de acceso a los datos y se suspende el proceso.

Por qué el navegador es el punto de control crítico

El navegador es donde realmente ocurre la mayor parte del uso de IA generativa, y también es una superficie de ataque intensamente atacada. Cyber Crucible lo ha comprobado de manera directa: desde el cuarto trimestre de 2023 en adelante, las respuestas automatizadas contra la actividad de Chrome, Edge y Chromium pasaron de cero a miles, con los CVE relacionados publicados posteriormente por Google y Microsoft. En una empresa, el patrón escaló de un puñado de ataques bloqueados a casi 4,000 en prácticamente todas las estaciones de trabajo.

¿Cómo sabe FortressAI que una herramienta de IA no ha sido manipulada?

Respuesta breve: Antes de conceder a cualquier herramienta de IA acceso a los datos, FortressAI evalúa si esa herramienta ha sido manipulada, comprobando la integridad de su proceso y de las bibliotecas cargadas mediante los mismos análisis de comportamiento de memoria en los que se apoya el resto de la plataforma. Una herramienta aprobada que ha sido comprometida no obtiene sus datos por el simple hecho de figurar en la lista de permitidas.

La aprobación no es lo mismo que la confianza

La mayoría de los marcos de gobernanza de IA se detienen en la lista de permitidas: ¿está autorizada esta aplicación? Eso es necesario, pero no suficiente, porque la identidad de una aplicación y su *integridad* son preguntas distintas.

`claude.exe`, un cliente de Copilot, una aplicación de escritorio de ChatGPT o un navegador que ejecuta Gemini pueden estar legítimamente aprobados, y aun así ser explotados. Un atacante que compromete un cliente de IA aprobado hereda el acceso a los datos que se le concedió a ese cliente. Una lista de permitidas por sí sola se lo entrega.

Por ello, FortressAI evalúa dos aspectos antes de conceder acceso a los datos:

1. **¿Está autorizada esta herramienta?** (evaluación y política por herramienta)
2. **¿Sigue siendo esta herramienta la misma?** (si el proceso o sus bibliotecas han sido manipulados)

Ambas condiciones deben cumplirse.

Qué significa "manipulada" en este contexto

La comprobación se basa en análisis de memoria, la misma capa de sensores fundamental que sustenta la protección de datos e identidad. Examina el estado del programa en ejecución y de las bibliotecas cargadas en busca de indicios de inyección, modificación en memoria o explotación.

Esto importa porque las técnicas de ataque modernas apuntan específicamente a procesos de confianza. El código inyectado en una aplicación aprobada hereda sus permisos y su reputación. Comprobar el archivo en disco no demuestra nada sobre en qué se ha convertido el proceso en memoria.

Qué ocurre cuando falla la comprobación

La respuesta sigue el mismo modelo escalonado utilizado en toda la plataforma:

- **Autorizada y sin manipular, dentro de la política** — el acceso continúa.
- **Autorizada y sin manipular, pero fuera de la política** — se bloquea, se redacta o se proporcionan datos falsos, según sus reglas. La herramienta sigue funcionando.
- **Manipulada o desconocida** — se rechaza o se suspende, según el motor de comportamiento y su configuración.

El efecto práctico: un cliente de IA comprometido se trata como un atacante, no como una aplicación aprobada que atraviesa un mal momento.

¿Puedo controlar qué herramientas de IA se permite usar en mi organización?

Respuesta breve: Sí. FortressAI ofrece evaluación por herramienta — control granular sobre exactamente qué herramientas de IA están autorizadas y cuáles están bloqueadas — y aplicación basada en la ubicación, de modo que la política pueda vincularse a dónde residen los datos sensibles en lugar de a aplicaciones individuales.

Dos controles complementarios

- **Evaluación por herramienta:** decida específicamente qué aplicaciones de IA (Gemini, Copilot y otras) están permitidas en su entorno.
- **Aplicación basada en la ubicación:** defina la protección en torno a las ubicaciones de datos que importan — manteniendo el código fuente en el repositorio, no en un chatbot — con asignaciones controladas por los usuarios.

El Protocolo del Semáforo

La política de FortressAI se expresa como un semáforo:

- **Rojo** — FortressAI bloquea que la herramienta de IA acceda a los datos asignados.
- **Amarillo** — el acceso se evalúa de forma condicional, incluido el manejo dinámico de datos sensibles. *(Las rutas condicionales que dependen de la aplicación de etiquetas de Purview/MIP y de la anonimización automática de PII/PDPL están en beta; consulte "¿Qué capacidades de FortressAI están disponibles de forma general y cuáles están en beta?")*
- **Verde** — el uso aprobado continúa con normalidad.

Esto permite a las organizaciones adoptar la IA de manera deliberada, en lugar de tener que elegir entre prohibiciones generales que terminan siendo eludidas y un acceso abierto que filtra datos.

¿Qué funcionalidades de FortressAI están disponibles con carácter general y cuáles se encuentran en fase beta?

Respuesta breve: La aplicación de controles a nivel de kernel — bloquear que datos sensibles lleguen a herramientas de IA, la autorización por herramienta y la política de datos basada en ubicación — está disponible con carácter general. La detección y aplicación de políticas sobre el uso de IA basado en navegador, la aplicación de etiquetas de confidencialidad de Microsoft Purview/MIP y la anonimización automática de PII/PDPL se encuentran actualmente en fase beta.

Disponibilidad general

- **Protección de datos a nivel de kernel** — se bloquea la salida de datos sensibles a través de ChatGPT, Copilot, Gemini o cualquier otra herramienta de IA, aplicada en la capa más profunda del sistema operativo.
- **Evaluación por herramienta** — control granular sobre qué herramientas de IA están exactamente autorizadas y cuáles están bloqueadas.
- **Aplicación basada en ubicación** — política vinculada a las ubicaciones de datos que importan, de modo que el código fuente permanezca en el repositorio en lugar de en un chatbot.

En fase beta

- **Detección y aplicación de políticas sobre el uso de IA basado en navegador** — supervisión y aplicación de políticas frente a la actividad de navegadores, páginas y extensiones.
- **Aplicación de etiquetas de confidencialidad de Microsoft Purview / MIP** — extiende la clasificación existente de Purview a una política por herramienta de IA, de modo que las organizaciones que ya clasifican sus datos no tengan que mantener un segundo esquema.
- **Anonimización automática de PII / PDPL** — filtrado dinámico que elimina la información de identificación personal de las instrucciones (prompts) que, por lo demás, están permitidas. Esto respalda el camino intermedio que la mayoría de las

organizaciones desean: permitir que las personas utilicen la IA de forma productiva, pero eliminando automáticamente los elementos sensibles en lugar de depender de que los empleados se autocensuren.

Las funcionalidades en fase beta son operativas y están en uso activo, pero deben evaluarse en su entorno antes de depender de ellas como control. Para conocer la disponibilidad actual de la fase beta e inscribirse, póngase en contacto con su representante de Cyber Crucible.

¿Cómo ayuda FortressAI con el cumplimiento normativo y las auditorías?

Respuesta breve: FortressAI previene la fuga en lugar de reportarla después de ocurrida, y proporciona registros claros de quién accedió a qué y dónde se detuvieron las fugas — evidencia que puede mostrar a un auditor.

Prevención como postura de cumplimiento

La mayor parte de la exposición al cumplimiento normativo derivada del uso de IA sigue el mismo patrón: los datos sensibles salen de la organización, y de ello se deriva la obligación de investigar, notificar o divulgar. Bloquear el acceso a nivel del kernel significa que el evento de divulgación nunca ocurre.

Demostrar el control

Los auditores y reguladores generalmente quieren dos cosas: evidencia de que existe un control, y evidencia de que funciona. FortressAI respalda ambas — una política que se aplica técnicamente a nivel del sistema operativo, y registros claros que muestran dónde se produjo la aplicación de dicha política.

Dónde se aplica esto

Este requisito aparece en muchos marcos normativos — HIPAA para información de pacientes, FERPA para registros estudiantiles, GDPR para datos personales, y obligaciones contractuales de confidencialidad. El control subyacente es el mismo en cada caso: los datos sensibles no deben salir del límite establecido, y usted debe poder demostrar que no lo hicieron.