FortressAI e Segurança de IA Generativa

Como o FortressAI governa o uso de IA generativa no nível do kernel - impedindo vazamentos de dados para ChatGPT, Copilot, Gemini e outras ferramentas de IA antes que aconteçam.

O que é o FortressAI?

Resposta curta: FortressAI é o produto de proteção de dados com IA generativa da Cyber Crucible. Ele aplica políticas na camada mais profunda do sistema operacional, verificando o acesso a dados no dispositivo e bloqueando, redigindo ou permitindo com segurança a informação antes que ela possa chegar ao ChatGPT, Copilot, Gemini ou qualquer outra ferramenta de IA.

O problema que resolve

As ferramentas públicas de IA estão transformando a forma como as pessoas trabalham e, ao mesmo tempo, abrindo um novo caminho para que os dados saiam de uma organização. Cada prompt carrega um risco. Os funcionários frequentemente compartilham informações sensíveis sem perceber que fizeram algo errado, e as ferramentas de segurança tradicionais nunca foram concebidas para isso.

Por que a aplicação em nível de kernel é importante aqui

O FortressAI opera na mesma profundidade que o restante da plataforma Cyber Crucible. Isso significa que ele protege contra *qualquer* ferramenta de IA — incluindo aquelas que ainda não existem — em vez de manter uma lista de aplicações específicas a controlar. Não é necessário compreender um novo produto de IA para impedir que dados sensíveis cheguem a ele.

Como a aplicação é local, o conteúdo dos prompts nunca é enviado a um serviço em nuvem para inspeção. A proteção não cria a exposição que se destina a evitar.

O FortressAI não é um mecanismo separado acoplado posteriormente. É uma camada de política sobre a mesma análise comportamental de memória do kernel que impulsiona a proteção de dados e identidade — o que permite que ele avalie não apenas se uma ferramenta de IA é aprovada, mas também se ela foi adulterada.

O que é Shadow AI, e quem dentro da minha organização está a criar o risco?

Resposta curta: Shadow AI refere-se a colaboradores que utilizam ferramentas de IA que o departamento de TI não aprovou ou desconhece. O risco não se limita ao pessoal de base — administradores privilegiados e contas comprometidas representam frequentemente a maior exposição.

Três fontes de exposição

- **Colaboradores** que utilizam ferramentas de IA não autorizadas para produtividade do dia a dia.
- **Administradores privilegiados**, cujo acesso de alto nível significa que tudo o que introduzem numa ferramenta de IA pode incluir dados corporativos altamente sensíveis.
- **Contas comprometidas**, em que um atacante com credenciais válidas utiliza ferramentas de IA como canal de exfiltração — tráfego que se assemelha a produtividade normal.

Isto já está a acontecer

A Cyber Crucible observou diretamente o acesso em massa a grandes volumes de dados de uma só vez, por colaboradores que utilizam ferramentas de IA para pesquisar e carregar informação. Este não é um risco futuro teórico a ser modelado; é um comportamento visível em implementações reais atualmente.

A terceira categoria merece atenção especial. Um atacante que utiliza ferramentas de IA para retirar dados é difícil de distinguir de um colaborador entusiasta — a menos que a aplicação de controlos ocorra ao nível do acesso aos dados, onde a intenção e a política podem ser avaliadas independentemente de quem está autenticado.

Por que a DLP, os firewalls de IA e os LLMs privados não conseguem impedir vazamentos de dados de IA?

Resposta curta: Cada abordagem trata de parte do problema a partir da camada errada. A DLP pode ser contornada, os firewalls de IA ficam na borda da rede e são cegos à atividade no dispositivo, e os LLMs privados são caros, sem fazer nada em relação aos funcionários que usam ferramentas públicas de qualquer forma.

Onde cada abordagem falha

- Os **LLMs privados** são caros de construir e operar, e não impedem que ninguém abra o ChatGPT em uma aba do navegador, independentemente disso.
- Os **firewalls de IA** inspecionam o tráfego de rede, o que os deixa cegos ao que acontece no próprio endpoint e à atividade interna (insider).
- A **DLP tradicional** foi projetada para arquivos e fluxos de e-mail, e é rotineiramente contornada por caminhos que nunca foi construída para monitorar.

O problema da lógica invertida

Várias abordagens nessa categoria tentam proteger os dados da IA inserindo *outra* IA para filtrar a saída. Isso adiciona custo e uma nova dependência sem tratar o ponto onde os dados realmente saem — o endpoint, no momento do acesso.

O FortressAI aplica a proteção no kernel, que está abaixo de todas essas camadas e se aplica independentemente de qual ferramenta, navegador ou extensão esteja envolvida.

Como o FortressAI protege dados em um navegador web?

> **Capacidade em beta.** A detecção e aplicação de políticas de uso de IA baseada em navegador está atualmente em beta. Ela é funcional e está em uso ativo, mas avalie-a em seu ambiente antes de depender dela como um controle.

Resposta curta: O FortressAI monitora a atividade do navegador a partir do kernel e a avalia em relação à política. Se o navegador, um site ou uma extensão tentar acessar dados fora da política, a permissão é revogada na camada do kernel — e, se o navegador apresentar sinais de exploração, todos os direitos de acesso a dados são revogados e o processo é suspenso.

A sequência

1. Um usuário abre um navegador web — Edge, Chrome ou Firefox.
2. O monitoramento em nível de kernel avalia o navegador e qualquer atividade de página ou extensão em relação à política do FortressAI atribuída.
3. Se o navegador, o site ou a extensão tentar acessar dados fora da política, a permissão é revogada na camada do kernel.
4. Se o navegador apresentar sinais de exploração, todos os direitos de acesso a dados são revogados e o processo é suspenso.

Por que o navegador é o ponto de controle crítico

O navegador é onde a maior parte do uso de IA generativa realmente acontece, e também é uma superfície de ataque fortemente visada. A Cyber Crucible tem observado isso diretamente: a partir do quarto trimestre de 2023, as respostas automatizadas contra atividades do Chrome, Edge e Chromium aumentaram de zero para milhares, com CVEs relacionados posteriormente publicados pelo Google e pela Microsoft. Em uma empresa, o padrão escalou de um punhado de ataques bloqueados para quase 4.000 em praticamente todas as estações de trabalho.

Como é que a FortressAI sabe que uma ferramenta de IA não foi comprometida?

Resposta breve: Antes de conceder a qualquer ferramenta de IA acesso a dados, a FortressAI avalia se essa ferramenta foi objeto de adulteração — verificando a integridade do seu processo e das bibliotecas carregadas através da mesma análise comportamental de memória em que assenta o resto da plataforma. Uma ferramenta aprovada que tenha sido comprometida não recebe os seus dados apenas por constar da lista de permissões.

Aprovação não é o mesmo que confiança

A maioria da governação de IA fica-se pela lista de permissões: esta aplicação está autorizada? Isso é necessário, mas não suficiente, porque a identidade de uma aplicação e a sua *integridade* são questões distintas.

O `claude.exe`, um cliente do Copilot, uma aplicação de ambiente de trabalho do ChatGPT ou um navegador a executar o Gemini podem todos ser legitimamente aprovados — e todos podem ser explorados. Um atacante que comprometa um cliente de IA aprovado herda todo o acesso a dados que foi concedido a esse cliente. Uma lista de permissões, por si só, entrega esse acesso.

Por isso, a FortressAI avalia dois aspetos antes de conceder acesso a dados:

1. **Esta ferramenta está autorizada?** (avaliação e política por ferramenta)
2. **Esta ferramenta continua a ser ela própria?** (o processo ou as suas bibliotecas foram adulterados)

Ambas as condições têm de se verificar.

O que significa "adulterado" neste contexto

A verificação recorre à análise de memória — a mesma camada de sensores fundamental que está na base da proteção de dados e de identidade. Esta análise examina o estado do programa em execução e das bibliotecas carregadas, à procura de sinais de injeção, modificação em memória ou exploração.

Isto é importante porque as técnicas de ataque modernas visam especificamente processos de confiança. Código injetado numa aplicação aprovada herda as suas permissões e a sua reputação. Verificar o ficheiro em disco nada prova sobre aquilo em que o processo se transformou em memória.

O que acontece quando a verificação falha

A resposta segue o mesmo modelo graduado utilizado em toda a plataforma:

- **Autorizado e não adulterado, dentro da política** — o acesso prossegue.
- **Autorizado e não adulterado, mas fora da política** — bloqueado, com dados ocultados ou substituídos por dados falsos, consoante as suas regras. A ferramenta continua a funcionar.
- **Adulterado ou desconhecido** — rejeitado ou suspenso, de acordo com o motor comportamental e as suas definições.

O efeito prático: um cliente de IA comprometido é tratado como um atacante, e não como uma aplicação aprovada que está simplesmente a ter um mau dia.

Posso controlar quais ferramentas de IA a minha organização tem permissão para usar?

Resposta curta: Sim. O FortressAI oferece avaliação por ferramenta — controlo granular sobre exatamente quais ferramentas de IA estão autorizadas e quais estão bloqueadas — e aplicação baseada na localização, para que a política possa estar associada ao local onde residem os dados sensíveis, em vez de a aplicações individuais.

Dois controlos complementares

- **Avaliação por ferramenta:** decida especificamente quais aplicações de IA (Gemini, Copilot, entre outras) são permitidas no seu ambiente.
- **Aplicação baseada na localização:** defina a proteção em torno dos locais de dados que importam — mantendo o código-fonte no repositório, e não num chatbot — com atribuições controladas pelos utilizadores.

O Protocolo de Semáforo (Traffic Light Protocol)

A política do FortressAI é expressa como um semáforo:

- **Vermelho** — o FortressAI bloqueia o acesso da ferramenta de IA aos dados atribuídos.
- **Amarelo** — o acesso é avaliado de forma condicional, incluindo o tratamento dinâmico de dados sensíveis. *(Os caminhos condicionais que dependem da aplicação de etiquetas Purview/MIP e da anonimização automática de PII/PDPL estão em fase beta — consulte "Quais funcionalidades do FortressAI estão disponíveis de forma geral e quais estão em beta?")*
- **Verde** — o uso aprovado prossegue normalmente.

Isto permite que as organizações adotem a IA de forma deliberada, em vez de terem de escolher entre proibições totais que acabam por ser contornadas e acessos abertos que provocam fugas de

dados.

Quais recursos do FortressAI estão disponíveis de forma geral e quais estão em beta?

Resposta curta: A aplicação de políticas em nível de kernel — bloqueio de dados sensíveis para que não cheguem a ferramentas de IA, autorização por ferramenta e política de dados baseada em localização — está disponível de forma geral. A detecção e aplicação de políticas de uso de IA baseada em navegador, a aplicação de rótulos de sensibilidade do Microsoft Purview/MIP e a anonimização automática de PII/PDPL estão atualmente em beta.

Disponível de forma geral

- **Proteção de dados em nível de kernel** — dados sensíveis são bloqueados de sair por meio do ChatGPT, Copilot, Gemini ou qualquer outra ferramenta de IA, com aplicação na camada mais profunda do sistema operacional.
- **Avaliação por ferramenta** — controle granular sobre exatamente quais ferramentas de IA são autorizadas e quais são bloqueadas.
- **Aplicação baseada em localização** — política vinculada às localizações de dados que importam, de modo que o código-fonte permaneça no repositório em vez de em um chatbot.

Em beta

- **Detecção e aplicação de uso de IA baseada em navegador** — monitoramento e aplicação de políticas em relação à atividade de navegador, página e extensões.
- **Aplicação de rótulos de sensibilidade do Microsoft Purview / MIP** — estendendo a classificação existente do Purview para uma política por ferramenta de IA, de modo que organizações que já classificam dados não precisem manter um segundo esquema.
- **Anonimização automática de PII / PDPL** — filtragem dinâmica que remove informações de identificação pessoal de prompts que, de outra forma, seriam permitidos. Isso oferece suporte ao caminho intermediário que a maioria das organizações deseja: permitir que as pessoas usem a IA de forma produtiva, mas removendo automaticamente os elementos sensíveis, em vez de depender que os funcionários se autocensurem.

Os recursos em beta são funcionais e estão em uso ativo, mas devem ser avaliados em seu ambiente antes de depender deles como um controle. Para informações atualizadas sobre

disponibilidade e inscrição no beta, entre em contato com seu representante da Cyber Crucible.

Como o FortressAI ajuda na conformidade e na auditoria?

Resposta breve: O FortressAI previne o vazamento em vez de relatá-lo posteriormente, e fornece registros claros de quem acessou o quê e onde os vazamentos foram interrompidos — evidências que você pode apresentar a um auditor.

Prevenção como postura de conformidade

A maior parte da exposição de conformidade decorrente do uso de IA segue o mesmo padrão: dados sensíveis saem da organização, e a obrigação de investigar, notificar ou divulgar segue-se a isso. Bloquear o acesso no nível do kernel significa que o evento de divulgação não ocorre.

Demonstrando controle

Auditores e reguladores geralmente querem duas coisas: evidência de que um controle existe e evidência de que ele funciona. O FortressAI oferece suporte a ambos — uma política que é tecnicamente aplicada no nível do sistema operacional, e registros claros mostrando onde a aplicação ocorreu.

Onde isso se aplica

O requisito aparece em muitas estruturas — HIPAA para informações de pacientes, FERPA para registros de estudantes, GDPR para dados pessoais, e obrigações contratuais de confidencialidade. O controle subjacente é o mesmo em cada caso: dados sensíveis não devem sair do limite estabelecido, e você deve ser capaz de provar que não saíram.