FortressAI & la sécurité de l'IA générative

Comment FortressAI régit l'utilisation de l'IA générative au niveau du noyau - en stoppant les fuites de données vers ChatGPT, Copilot, Gemini et d'autres outils d'IA avant qu'elles ne se produisent.

Qu'est-ce que FortressAI ?

Réponse courte : FortressAI est le produit de protection des données par IA générative de Cyber Crucible. Il applique la politique au niveau le plus profond du système d'exploitation, en vérifiant l'accès aux données sur l'appareil et en bloquant, en masquant ou en autorisant en toute sécurité les informations avant qu'elles puissent atteindre ChatGPT, Copilot, Gemini ou tout autre outil d'IA.

Le problème qu'il résout

Les outils d'IA publics transforment la façon dont les gens travaillent, tout en ouvrant simultanément une nouvelle voie par laquelle les données peuvent quitter une organisation. Chaque invite comporte un risque. Les employés partagent fréquemment des informations sensibles sans se rendre compte qu'ils ont commis une erreur, et les outils de sécurité traditionnels n'ont jamais été conçus pour cela.

Pourquoi l'application au niveau du noyau est importante ici

FortressAI fonctionne à la même profondeur que le reste de la plateforme Cyber Crucible. Cela signifie qu'il protège contre *tout* outil d'IA — y compris ceux qui n'existent pas encore — plutôt que de maintenir une liste d'applications spécifiques à surveiller. Il n'a pas besoin de comprendre un nouveau produit d'IA pour empêcher les données sensibles de l'atteindre.

Comme l'application des règles est locale, le contenu des invites n'est jamais transmis à un service cloud pour inspection. La protection ne crée pas l'exposition qu'elle est censée prévenir.

FortressAI n'est pas un moteur distinct ajouté après coup. Il s'agit d'une couche de politique s'appuyant sur les mêmes analyses comportementales de la mémoire du noyau qui alimentent la protection des données et des identités — c'est ce qui lui permet de juger non seulement si un outil d'IA est approuvé, mais aussi s'il a été altéré.

Qu'est-ce que le Shadow AI, et qui, au sein de mon organisation, crée le risque ?

Réponse courte : Le Shadow AI désigne les employés qui utilisent des outils d'IA que le service informatique n'a pas approuvés ou dont il ignore l'existence. Le risque ne se limite pas au personnel non hiérarchique — les administrateurs privilégiés et les comptes compromis constituent souvent une exposition plus importante.

Trois sources d'exposition

- **Les employés** utilisant des outils d'IA non autorisés pour leur productivité quotidienne.
- **Les administrateurs privilégiés**, dont l'accès de haut niveau signifie que tout ce qu'ils transmettent à un outil d'IA peut inclure des données d'entreprise hautement sensibles.
- **Les comptes compromis**, où un attaquant disposant d'identifiants valides utilise les outils d'IA comme canal d'exfiltration — un trafic qui ressemble à une activité de productivité ordinaire.

Cela se produit déjà

Cyber Crucible a directement observé des accès massifs à de gros volumes de données en une seule fois, par des employés utilisant des outils d'IA pour rechercher et téléverser des informations. Il ne s'agit pas d'un risque futur théorique en cours de modélisation ; c'est un comportement visible dans des déploiements réels dès aujourd'hui.

La troisième catégorie mérite une attention particulière. Un attaquant utilisant des outils d'IA pour exfiltrer des données est difficile à distinguer d'un employé enthousiaste — à moins que l'application des règles n'intervienne au niveau de l'accès aux données, où l'intention et la politique peuvent être évaluées indépendamment de l'identité de la personne connectée.

Pourquoi la DLP, les pare-feux IA et les LLM privés ne parviennent-ils pas à empêcher les fuites de données liées à l'IA ?

Réponse courte : Chacune de ces solutions traite une partie du problème depuis la mauvaise couche. La DLP peut être contournée, les pare-feux IA se situent en périphérie du réseau et sont aveugles à l'activité sur l'appareil, et les LLM privés coûtent cher tout en ne faisant rien contre les employés qui utilisent malgré tout des outils publics.

Où chaque approche est insuffisante

- Les **LLM privés** sont coûteux à construire et à exploiter, et n'empêchent en rien quiconque d'ouvrir ChatGPT dans un onglet de navigateur.
- Les **pare-feux IA** inspectent le trafic réseau, ce qui les rend aveugles à ce qui se passe sur le point de terminaison lui-même et à l'activité interne (insider).
- La **DLP traditionnelle** a été conçue pour les fichiers et les flux de messagerie, et est régulièrement contournée par des voies qu'elle n'a jamais été conçue pour surveiller.

Le problème de la logique inversée

Plusieurs approches de cette catégorie tentent de protéger les données contre l'IA en insérant une *autre* IA pour filtrer la sortie. Cela ajoute un coût et une nouvelle dépendance sans traiter l'endroit où les données sortent réellement — le point de terminaison, au moment de l'accès.

FortressAI applique ses règles au niveau du noyau, en dessous de toutes ces couches, et s'applique quel que soit l'outil, le navigateur ou l'extension concerné.

Comment FortressAI protège-t-il les données dans un navigateur web ?

> **Fonctionnalité en version bêta.** La détection et l'application des politiques d'utilisation de l'IA dans le navigateur sont actuellement en version bêta. Cette fonctionnalité est opérationnelle et activement utilisée, mais évaluez-la dans votre environnement avant de vous y fier comme contrôle.

Réponse courte : FortressAI surveille l'activité du navigateur depuis le noyau et l'évalue par rapport à la politique définie. Si le navigateur, un site web ou une extension tente d'accéder à des données en dehors de la politique, l'autorisation est révoquée au niveau du noyau — et si le navigateur montre des signes d'exploitation, tous les droits d'accès aux données sont révoqués et le processus est suspendu.

Le déroulement

1. Un utilisateur ouvre un navigateur web — Edge, Chrome ou Firefox.
2. La surveillance au niveau du noyau évalue le navigateur ainsi que toute activité de page ou d'extension par rapport à la politique FortressAI assignée.
3. Si le navigateur, le site ou l'extension tente d'accéder à des données en dehors de la politique, l'autorisation est révoquée au niveau du noyau.
4. Si le navigateur montre des signes d'exploitation, tous les droits d'accès aux données sont révoqués et le processus est suspendu.

Pourquoi le navigateur est le point de contrôle critique

Le navigateur est l'endroit où se déroule la majeure partie de l'utilisation réelle de l'IA générative, et c'est également une surface d'attaque fortement ciblée. Cyber Crucible en a été directement témoin : à partir du quatrième trimestre 2023, les réponses automatisées contre l'activité de Chrome, Edge et Chromium sont passées de zéro à plusieurs milliers, avec des CVE associées publiées par la suite par Google et Microsoft. Dans une entreprise, le schéma a évolué d'une

poignée d'attaques bloquées à près de 4 000, touchant presque tous les postes de travail.

Comment FortressAI sait-il qu'un outil d'IA n'a pas été altéré ?

Réponse courte : Avant d'accorder à un outil d'IA l'accès à des données, FortressAI évalue si cet outil a été altéré — en vérifiant l'intégrité de son processus et des bibliothèques chargées grâce aux mêmes analyses comportementales de la mémoire sur lesquelles repose le reste de la plateforme. Un outil approuvé qui a été compromis n'obtient pas vos données simplement parce qu'il figure sur la liste autorisée.

L'approbation n'est pas synonyme de confiance

La plupart des dispositifs de gouvernance de l'IA s'arrêtent à la liste d'autorisation : cette application est-elle permise ? C'est nécessaire, mais insuffisant, car l'identité d'une application et son *intégrité* sont deux questions distinctes.

`claude.exe`, un client Copilot, une application de bureau ChatGPT, ou un navigateur exécutant Gemini peuvent tous être légitimement approuvés — et tous être exploités. Un attaquant qui compromet un client IA approuvé hérite de tout accès aux données accordé à ce client. Une liste d'autorisation seule le lui remet.

FortressAI évalue donc deux éléments avant l'accès aux données :

1. **Cet outil est-il autorisé ?** (évaluation et politique propres à chaque outil)
2. **Cet outil est-il toujours lui-même ?** (le processus ou ses bibliothèques ont-ils été altérés)

Les deux conditions doivent être remplies.

Ce que signifie « altéré » dans ce contexte

Le contrôle s'appuie sur les analyses de la mémoire — la même couche de capteurs fondamentale qui sous-tend la protection des données et de l'identité. Il examine l'état du programme en cours d'exécution et de ses bibliothèques chargées à la recherche de signes d'injection, de modification en mémoire ou d'exploitation.

Cela importe car les techniques d'attaque modernes visent spécifiquement les processus de confiance. Un code injecté dans une application approuvée hérite de ses permissions et de sa réputation. Vérifier le fichier sur le disque ne prouve rien quant à ce que le processus est devenu en mémoire.

Que se passe-t-il en cas d'échec du contrôle

La réponse suit le même modèle graduel appliqué dans l'ensemble de la plateforme :

- **Autorisé et non altéré, conforme à la politique** — l'accès se poursuit.
- **Autorisé et non altéré, mais dépassant la politique** — bloqué, expurgé, ou alimenté en fausses données, selon vos règles. L'outil continue de fonctionner.
- **Altéré ou inconnu** — rejeté ou suspendu, selon le moteur comportemental et vos paramètres.

L'effet pratique : un client IA compromis est traité comme un attaquant, et non comme une application approuvée traversant un moment difficile.

Puis-je contrôler quels outils d'IA mon organisation est autorisée à utiliser ?

Réponse courte : Oui. FortressAI fournit une évaluation par outil — un contrôle granulaire permettant de déterminer exactement quels outils d'IA sont autorisés et lesquels sont bloqués — ainsi qu'une application basée sur l'emplacement, de sorte que la politique puisse être liée à l'endroit où résident les données sensibles plutôt qu'à des applications individuelles.

Deux contrôles complémentaires

- **Évaluation par outil :** décidez précisément quelles applications d'IA (Gemini, Copilot, et autres) sont autorisées dans votre environnement.
- **Application basée sur l'emplacement :** définissez une protection autour des emplacements de données qui comptent — en conservant le code source dans le référentiel, et non dans un chatbot — avec des affectations contrôlées par les utilisateurs.

Le protocole des feux tricolores (Traffic Light Protocol)

La politique de FortressAI s'exprime sous la forme d'un feu tricolore :

- **Rouge** — FortressAI bloque l'accès de l'outil d'IA aux données assignées.
- **Jaune** — l'accès est évalué de manière conditionnelle, y compris le traitement dynamique des données sensibles. *(Les chemins conditionnels qui dépendent de l'application des étiquettes Purview/MIP et de l'anonymisation automatique des PII/PDPL sont en version bêta — voir « Quelles fonctionnalités de FortressAI sont généralement disponibles, et lesquelles sont en version bêta ? »)*
- **Vert** — l'utilisation approuvée se poursuit normalement.

Cela permet aux organisations d'adopter l'IA de manière délibérée, plutôt que de choisir entre des interdictions générales que l'on contourne et un accès ouvert qui provoque des fuites de données.

Quelles fonctionnalités de FortressAI sont généralement disponibles, et lesquelles sont en version bêta ?

Réponse courte : L'application au niveau du noyau — le blocage des données sensibles empêchant qu'elles n'atteignent les outils d'IA, l'autorisation par outil et la politique de données basée sur la localisation — est généralement disponible. La détection et l'application de l'utilisation de l'IA basée sur le navigateur, l'application des étiquettes de sensibilité Microsoft Purview/MIP et l'anonymisation automatique des données PII/PDPL sont actuellement en version bêta.

Généralement disponible

- **Protection des données au niveau du noyau** — les données sensibles sont bloquées lorsqu'elles tentent de sortir via ChatGPT, Copilot, Gemini ou tout autre outil d'IA, l'application étant assurée au niveau le plus profond du système d'exploitation.
- **Évaluation par outil** — contrôle granulaire permettant de déterminer précisément quels outils d'IA sont autorisés et lesquels sont bloqués.
- **Application basée sur la localisation** — la politique est liée aux emplacements de données qui comptent, de sorte que le code source reste dans le référentiel plutôt que dans un chatbot.

En version bêta

- **Détection et application de l'utilisation de l'IA basée sur le navigateur** — surveillance et application de la politique concernant l'activité du navigateur, des pages et des extensions.
- **Application des étiquettes de sensibilité Microsoft Purview / MIP** — extension de la classification Purview existante vers une politique par outil d'IA, afin que les organisations qui classifient déjà leurs données n'aient pas à maintenir un second système.
- **Anonymisation automatique des données PII / PDPL** — filtrage dynamique qui supprime les informations personnellement identifiables des invites (prompts) par ailleurs autorisées. Cela permet la voie intermédiaire que la plupart des organisations recherchent

: laisser les utilisateurs se servir de l'IA de manière productive, tout en supprimant automatiquement les éléments sensibles plutôt que de compter sur l'autocensure des employés.

Les fonctionnalités en version bêta sont opérationnelles et activement utilisées, mais doivent être évaluées dans votre environnement avant que vous n'en dépendiez pour un contrôle. Pour connaître la disponibilité actuelle des versions bêta et les modalités d'inscription, contactez votre représentant Cyber Crucible.

Comment FortressAI aide-t-il en matière de conformité et d'audit ?

Réponse courte : FortressAI empêche la fuite plutôt que de la signaler après coup, et fournit des registres clairs indiquant qui a accédé à quoi et où les fuites ont été stoppées — des preuves que vous pouvez présenter à un auditeur.

La prévention comme posture de conformité

La plupart des expositions liées à la conformité découlant de l'utilisation de l'IA suivent le même schéma : des données sensibles quittent l'organisation, et l'obligation d'enquêter, de notifier ou de divulguer s'ensuit. Bloquer l'accès au niveau du noyau signifie que l'événement de divulgation ne se produit pas.

Démontrer le contrôle

Les auditeurs et les régulateurs veulent généralement deux choses : la preuve qu'un contrôle existe, et la preuve qu'il fonctionne. FortressAI répond aux deux exigences — une politique techniquement appliquée au niveau du système d'exploitation, et des registres clairs indiquant où l'application a eu lieu.

Où cela s'applique

Cette exigence apparaît dans de nombreux cadres réglementaires — HIPAA pour les informations relatives aux patients, FERPA pour les dossiers des élèves, GDPR pour les données personnelles, et les obligations contractuelles de confidentialité. Le contrôle sous-jacent est le même dans chaque cas : les données sensibles ne doivent pas franchir la limite, et vous devez être en mesure de prouver qu'elles ne l'ont pas fait.