Le défi de la sécurité de l'IA : Guide de déploiement du Zero Trust à l'intention des équipes de sécurité

Un guide pratique pour sécuriser l'ensemble du cycle de vie de l'IA, de l'identification de l'IA fantôme à la protection des agents d'IA en production, grâce aux principes du Zero Trust, à une feuille de route de déploiement en 30, 60 et 90 jours et à des indicateurs de performance (KPI) décisionnels.

E-BOOK

Introduction

Les équipes de sécurité le savent : l'IA leur pose un problème. Mais elles en ignorent souvent l'ampleur.

Quelles sont les applications d'intelligence artificielle (IA) actives au sein de votre environnement actuel ? Quels sont les modèles interrogés par vos développeurs via les plug-ins de leur environnement de développement intégré (IDE) et leurs outils de ligne de commande (CLI) ? Quelles sont les données insérées dans les prompts ? Contiennent-elles du code source, des données personnelles de clients ou des contenus dont la divulgation est réglementée ? Quels sont les agents qui agissent de manière autonome pour le compte des utilisateurs et à quels systèmes peuvent-ils accéder ?

Le véritable risque réside dans l'écart entre ce que fait l'IA dans votre environnement et ce que vous êtes en mesure de voir et de piloter. Il ne relève pas de scénarios d'attaque théoriques, mais de l'utilisation quotidienne, non approuvée mais de bonne foi, d'outils d'IA non validés par les équipes de sécurité, souvent parce que personne n'a pensé à le faire. L'informatique fantôme et les fonctionnalités d'IA intégrées aux plateformes SaaS déjà approuvées sont les principales sources de cette zone d'ombre. Un outil validé par le service informatique pour un usage donné peut, suite à une mise à jour, intégrer un assistant IA qui traite désormais des données internes sans avoir fait l'objet de la moindre évaluation de sécurité.

Les données et statistiques présentées dans ce guide sont issues du rapport ThreatLabz 2026 sur la sécurité de l'IA, publié par Zscaler en 2026. Elles résultent de l'analyse de 989,3 milliards de transactions IA/AA observées sur la plateforme Zscaler Zero Trust Exchange entre janvier et décembre 2025. Sauf indication contraire, tous les chiffres correspondent à cette période d'étude.

Le Zero Trust impose des principes clairs et de bon sens :

Aucune application, aucun utilisateur, aucun prompt, ni aucun connecteur ne peut être considéré par défaut comme étant de confiance. Des contrôles d'accès fondés sur l'identité et le contexte doivent être appliqués à tous les niveaux. Les évaluations doivent être permanentes. Le périmètre de dommages suite à un incident doit être minimisé. Les principes du Zero Trust doivent être mis en oeuvre dans le bon ordre. La visibilité doit précéder la mise en oeuvre des politiques. Les contrôles d'accès doivent être stabilisés avant d'être automatisés. Déployés dans le mauvais ordre, ces principes risquent d'entraver des activités légitimes ou de passer à côté de risques réels.

83 %	**18 033 To**	**39 %**	**93 %**
Croissance annuelle des transactions IA/AA en entreprise , à l'échelle de l'infrastructure	de données transférées via des applications d'IA en 2025 — plus que ce que la plupart des programmes de sécurité peuvent suivre	des transactions d'IA sont activement bloquées — la gouvernance est déjà sous pression.	Croissance annuelle du volume des données IA transférées — un rythme que les programmes de sécurité ne peuvent tenir

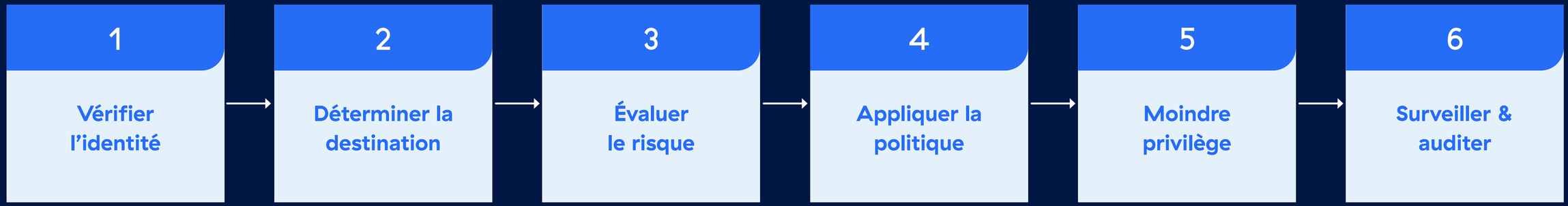
Source : Rapport ThreatLabz 2026 sur la sécurité de l'IA, Zscaler (p. 5—6)

Les principaux contenus de ce guide

1. Les sept principes du Zero Trust appliqués aux systèmes, aux agents et aux flux de données liés à l'IA
2. Une feuille de route de déploiement sur 30, 60 et 90 jours, structurée en des étapes interdépendantes
3. Les principaux risques émergents liés à l'IA : IA fantôme, injection de prompts, systèmes agentiques, trafic MCP (Model Context Protocol) et A2A (agent-to-agent), IA embarquée dans le SaaS et posture de sécurité des infrastructures d'IA
4. Un modèle d'indicateurs clés de performance (KPI) pour rendre compte de l'avancement du programme de sécurité de l'IA

Les principes du Zero Trust appliqués à l’IA

Les sept principes ci-dessous constituent un cadre décisionnel pour l’ensemble de la feuille de route. Chacun d’eux permet d’évaluer les fonctionnalités à déployer. Ils sont présentés dans l’ordre de déploiement sur le terrain.



À appliquer dans cet ordre. Chaque principe s’appuie sur le précédent.

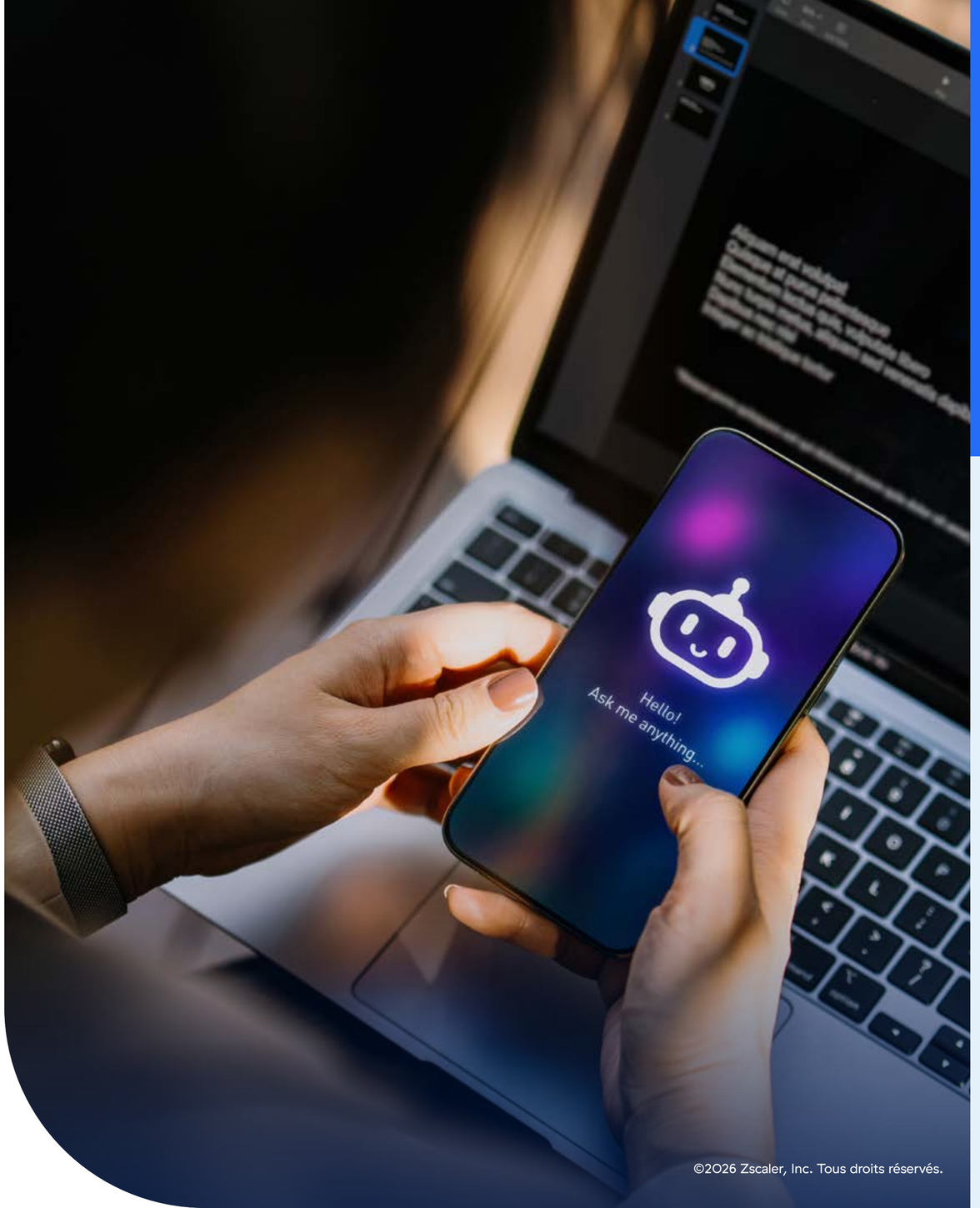
Principe n° 1 :
VÉRIFIER L'IDENTITÉ DES UTILISATEURS, DES DÉVELOPPEURS ET DES SERVICES

La gestion des accès à l'IA commence par les identités. Chaque utilisateur accédant à un outil SaaS d'IA générative (GenAI), chaque développeur interrogeant un modèle via un plug-in de son IDE et chaque compte de service se connectant à une composante d'une infrastructure d'IA doit être authentifié par l'intermédiaire d'un plan unique et cohérent de contrôle des identités. L'authentification unique (SSO) et l'authentification multifacteur (MFA) sont essentielles, mais elles ne suffisent pas à elles seules.

Les utilisateurs et les développeurs présentent des profils de risque différents et doivent donc faire l'objet de politiques distinctes. Les risques liés à l'exposition des données ne sont pas les mêmes pour un analyste marketing qui utilise ChatGPT pour rédiger du contenu que pour un ingénieur logiciel dont l'IDE transmet des demandes de génération de code à GitHub Copilot via un serveur MCP connecté à des référentiels de code internes. L'authentification des agents d'IA à un service constitue un nouveau cas d'usage auquel la plupart des entreprises ne sont pas encore préparées.

Principales mesures de contrôle

- Application du SSO et de la MFA aux services SaaS avec IA et aux outils d'IA destinés aux développeurs.
- Politiques d'identité distinctes pour les utilisateurs d'IA générative et les outils d'IA destinés aux développeurs (IDE, CLI, intégrations MCP)
- **Contrôles des comptes de service liés à des agents d'IA :** application du moindre privilège, identifiants éphémères et journalisation des audits
- **Posture de sécurité des terminaux comme indicateur d'identité :** les terminaux gérés bénéficient de droits d'accès différents de ceux des équipements personnels (BYOD)



Principe n° 2 :
DÉTERMINER LES DESTINATIONS POUR L'IA AUTORISÉE ET NON AUTORISÉE

Toutes les applications d'IA ne présentent pas le même profil de risque et une politique consistant à tout bloquer n'est ni réaliste sur le plan opérationnel ni efficace du point de vue de la sécurité. **L'objectif est de mettre en place une politique de destination structurée en plusieurs niveaux :** les outils autorisés bénéficient d'un accès complet avec inspection inline, les outils tolérés bénéficient d'un accès surveillé et soumis à des garde-fous, tandis que les outils non autorisés ou présentant un risque élevé sont bloqués ou isolés. La liste de ces périmètres liés à l'IA est bien plus vaste que ne l'imaginent la plupart des entreprises : elle a été multipliée par quatre en seulement un an, à mesure que de nouvelles fonctionnalités d'IA ont été intégrées aux plateformes existantes.

Principales mesures de contrôle

- Politiques de destination en plusieurs niveaux permettant d'autoriser, d'avertir, d'isoler ou de bloquer, selon l'application utilisée et le groupe d'utilisateurs
- Découverte dynamique des applications afin d'identifier les nouveaux outils d'IA dès leur apparition dans le trafic
- Politiques distinctes pour la GenAI en mode SaaS, l'IA intégrée aux applications SaaS et les outils d'IA destinés aux développeurs.
- Processus de validation des demandes d'utilisation de nouveaux outils d'IA, pour offrir une solution aux utilisateurs plutôt que de les amener à contourner les mesures en place

3 400

APPLICATIONS D'IA DISTINCTES RECENSÉES DANS LE TRAFIC DES ENTREPRISES, UN NOMBRE MULTIPLIÉ PAR QUATRE EN UN AN ET QUI CONTINUE D'AUGMENTER CHAQUE MOIS

Rapport ThreatLabz 2026 sur la sécurité l'IA, Zscaler (p. 5)

Principe n° 3 :
ÉVALUER LE RISQUE EN FONCTION DU CONTEXTE

Les décisions liées aux accès à l'IA ne doivent pas être binaires. Pour un même utilisateur accédant au même outil d'IA, le niveau de risque varie selon le contexte : dispositif utilisé, sensibilité des données et type d'action. Une évaluation efficace des risques suppose de définir en amont les catégories de données sensibles (« red data ») : code source, données personnelles, données de cartes de paiement (PCI), informations médicales protégées, résultats financiers non publiés ou encore informations confidentielles relatives aux opérations de fusion–acquisition (M&A). Les principales catégories de violations observées dans le trafic des entreprises reflètent précisément les types d'informations soumis par les collaborateurs : données nominatives et identifiants, code source et dossiers médicaux.

Principales mesures de contrôle

- Définition explicite des catégories de « red data » pour le code source, données personnelles/PCI/médicales et les projets non divulgués publiquement
- Évaluation du risque selon le groupe d'utilisateurs, le service, la posture de sécurité du terminal et la sensibilité des données
- Adaptation dynamique des politiques en fonction du contexte, et non sur la seule base de règles statiques d'autorisation ou de blocage
- Traitement différencié des interactions avec l'IA selon qu'elles comportent un téléversement de fichier ou uniquement des prompts textuels

410 M

VIOLATIONS DE LA PRÉVENTION DES PERTES DE DONNÉES (DLP) SUR CHATGPT UNIQUEMENT : DIVULGATION DE NOMS, DE NUMÉROS DE SÉCURITÉ SOCIALE, DE CODE SOURCE, DE DONNÉES MÉDICALES...

Rapport ThreatLabz 2026 sur la sécurité l'IA, Zscaler (pp. 6, 15)

Principe n° 4 :
APPLIQUER LES POLITIQUES EN MODE INLINE, PAR SESSION

Les journaux ne font que retracer les événements. Lorsqu'un log indique qu'un développeur a soumis du code source propriétaire à un modèle IA non autorisé, les données ont déjà quitté l'entreprise. La plupart des entreprises aggravent cette situation en déployant la DLP uniquement pour les téléversements de fichiers vers les outils d'IA, tandis que le contenu des prompts n'est pas contrôlé. Il suffit qu'un développeur saisisse directement le code source dans une fenêtre de prompt pour contourner les mesures de contrôle en place, sans alerte. La croissance du volume des transferts de données illustre l'ampleur des informations qui transitent par ces canaux non protégés.

Fonctionnalités clés

- Inspection DLP inline des prompts, des réponses et des téléversements de fichiers, et pas seulement de l'URL ou des catégories d'URL
- Possibilité d'autoriser, d'avertir, de masquer, de bloquer ou d'isoler au niveau des points d'application des politiques
- Classification des prompts pour les principales applications d'IA générative, avec détection contextuelle
- Modération des contenus pour les interactions hors sujet, soumises à des restrictions ou présentant un risque concurrentiel
- Isolation du navigateur pour les outils d'IA autorisés mais à risque : accès autorisé mais sans confiance totale

93 %

CROISSANCE ANNUELLE DU VOLUME DES TRANSFERTS DE DONNÉES LIÉS À L'IA EN ENTREPRISE. UN TEL RYTHME DÉPASSE LES CAPACITÉS DE TRACKING ET DE GOUVERNANCE DE LA PLUPART DES PROGRAMMES DE SÉCURITÉ.

Rapport ThreatLabz 2026 sur la sécurité l'IA, Zscaler (p. 5, 14)

Principe n° 5 :
MINIMISER LES ACCÈS ET L'EXPOSITION AUX RISQUES

Les systèmes d'IA créent des voies d'accès qui échappent souvent aux modèles traditionnels de gestion des identités et de sécurité réseau.Un pipeline RAG (Retrieval-Augmented Generation) peut accorder à un modèle IA l'accès à un corpus documentaire.Un agent peut se connecter à des outils internes, des jeux de données, des bases vectorielles ou des serveurs MCP. Ces autorisations sont tout aussi concrètes que celles accordées aux utilisateurs, même si elles n'apparaissent pas dans les audits d'accès traditionnels.

Réduire les accès et limiter l'exposition consiste à restreindre les ressources auxquelles les systèmes d'IA peuvent accéder et les communications qu'ils peuvent établir. Les agents, pipelines de données, composants RAG et sources de données back-end ne devraient communiquer qu'au travers de chemins explicitement autorisés. Sans segmentation selon le principe du moindre privilège, un agent disposant d'autorisations excessives ou un composant d'IA compromis peut favoriser le déplacement latéral de menaces.

Fonctionnalités clés

- Définition des accès selon le rôle pour les applications d'IA, les agents, les connecteurs des modèles, les jeux de données, les bases vectorielles et les composants de l'infrastructure d'IA
- Audit régulier des droits d'accès et réduction des accès à privilèges excessifs aux jeux de données sensibles et aux systèmes connectés à l'IA
- Permissions minimales pour les agents d'IA, pour définir précisément ce qu'ils peuvent atteindre et interdire tout le reste
- Segmentation des services d'IA, des plateformes d'agents, des pipelines de données, des composants RAG et des sources de données back-end
- Restreindre les accès entre les composants d'IA aux seuls chemins explicitement autorisés
- Inspection et journalisation des échanges spécifiques au protocole MCP et A2A

Limiter le périmètre d'impact d'une menace sur chaque chemin d'accès à l'IA

Considérez chaque connecteur d'IA, chaque autorisation accordée à un agent, chaque pipeline RAG, chaque base vectorielle, chaque serveur MCP et chaque connexion entre composants comme un point d'exposition aux risques. Définissez précisément les ressources auxquelles l'IA peut accéder, limitez ses communications, inspectez le trafic et consignez l'activité.

Principe n° 6 :
SURVEILLANCE CONTINUE ET TRAÇABILITÉ

La vérification continue est un mode de fonctionnement, et non une étape à franchir avant de passer à autre chose. Pour les systèmes d'IA, cette pratique implique de journaliser les prompts et les réponses, de déclencher des alertes en temps réel en cas de violation des politiques, de disposer de tableaux de bord sur les usages et de mettre en oeuvre un processus d'investigation qui évite d'avoir à recouper manuellement les journaux provenant de différents outils. **La pression réglementaire est bien réelle :** le règlement européen sur l'intelligence artificielle (EU AI Act), le cadre de gestion des risques liés à l'intelligence artificielle du NIST (NIST AI RMF) et les nouveaux cadres de gouvernance de l'IA imposent tous des contrôles documentés et démontrables.

Fonctionnalités clés

- Journalisation des prompts et des réponses pour les interactions IA
- Tableaux de bord d'utilisation par utilisateur, département métier et application d'IA, enrichis d'indicateurs de risque
- Alertes et processus d'investigation pour les violations de la politique et les anomalies liées à l'IA
- Reporting à des fins d'audit : IA utilisées, contrôles appliqués et violations constatées
- Mapping de la gouvernance par rapport aux exigences du NIST AI RMF, de l'EU AI Act et d'autres référentiels applicables

1	2	3	4	5
Réactif Aucun inventaire Blocage au cas par cas	**Visible** Empreinte IA identifiée Politiques de référence	**Gouverné** Politiques fondées sur les rôles Workflows de gouvernance	**Automatisé** Coaching automatisé Conformité permanente	**Continu** Red teaming + Garde-fous en environnement de production

Maturité du programme →

Jours 1 à 30

Établir la visibilité

Vous ne pouvez sécuriser une IA sur laquelle vous n'avez aucune visibilité. Les 30 premiers jours sont consacrés à établir un inventaire et à la mise en place des premières mesures de contrôle, afin de prendre des décisions éclairées plutôt que de tout bloquer d'emblée.

JOURS 1 À 30 Établir la visibilité	JOURS 31 À 60 Rendre les politiques plus matures	JOURS 61 À 90 Automatiser et valider
• Identifier les applications d'IA • Définir les catégories de « red data » • Politiques d'accès de référence • Premier tableau de bord décisionnel	• Politiques d'accès fondées sur les rôles • Processus de gouvernance en place • Déploiement de l'IA-SPM • Protections pour les IA privées	• Coaching automatisé des utilisateurs • Conformité permanente • Red teaming régulier • Garde-fous en boucle fermée

Feuille de route de déploiement sur 30, 60 et 90 jours

Checklist des 30 premiers jours

- **Mettre en place un examen périodique des outils d'IA :** nouveaux outils détectés dans le trafic, demandes de dérogation reçues et changement dans les politiques
- Créer le premier tableau de bord de sécurité de l'IA destiné aux décideurs : parc des applications d'IA (autorisées et fantômes), événements DLP par catégorie, actions des politiques et tendances
- Définir des playbooks de réponse aux incidents pour les principaux événements de sécurité liés à l'IA : données sensibles dans les prompts, accès à des destinations bloquées et téléversements de fichiers volumineux vers des outils d'IA.
- **Centraliser les journaux de sécurité liés à l'IA dans un seul tableau de bord :** activité des utilisateurs, utilisation des applications, événements liés aux prompts et aux réponses, application des politiques et violations de règles.

Déployer les mesures de contrôle

- **Passer en revue et affiner les politiques de DLP à partir des indicateurs recueillis au cours des deux premières semaines** , ceci afin de réduire les faux positifs et de mieux détecter les données exposées
- **Modérer le contenu pour faire respecter les règles d'utilisation :** interactions hors sujet, sujets soumis à restrictions, contenus sensibles sur le plan concurrentiel et réponses inappropriées
- **Ajouter l'isolation du navigateur pour les outils d'IA de la catégorie « autorisés mais à risque » :** permettre l'accès à l'outil tout en limitant les téléversements, les téléchargements et les opérations de copier-coller à partir du navigateur.
- **Étendre la couverture DLP à l'ensemble des catégories des « red data » :** code source (langages et signatures supplémentaires), données personnelles (au-delà des formats les plus courants), données de titulaires de cartes de paiement, données médicales et contenus métier confidentiels

Renforcer la protection des données

- **Publier un guide de l'utilisation de l'IA fondé sur des codes de couleur :** vert (autorisé), jaune (utilisation soumise à restrictions), rouge (bloqué, avec justification). Indiquer également la procédure de demande de dérogation
- Activer la DLP inline pour les prompts, en commençant par des règles de détection fiables pour les catégories de « red data » : signatures de code source, formats d'identifiants et schémas de données personnelles.
- **Restreindre les actions les plus risquées sur les outils non autorisés :** téléversements, copier-coller de gros volumes et téléchargement des contenus générés par l'IA vers des espaces de stockage non contrôlés
- **Appliquer la politique de destination :** alerte ou blocage selon le groupe d'utilisateurs et l'application pour les destinations IA non autorisées ou présentant un risque élevé

Appliquer les contrôles de base

- **Établir un inventaire des risques liés à l'IA :** identifier les applications qui accèdent à chaque type de données et les groupes d'utilisateurs qui recourent le plus à l'IA
- **Définir explicitement les catégories de « red data » :** code source, données personnelles, données PCI, informations de santé, secrets et identifiants, données financières non publiées et données sensibles relatives aux opérations de fusion-acquisition (M&A) ou aux projets stratégiques

Définir les données à protéger

- Activer la visibilité sur les prompts des applications d'IA les plus utilisées afin de comprendre les données échangées via les prompts et les réponses
- Identifier les outils d'IA intégrés avec des systèmes internes et les types de données auxquels ils accèdent
- **Identifier l'IA fantôme :** outils activement utilisés sans être officiellement approuvés.
- Lancer la fonction d'identification des applications d'IA afin d'inventorier les services SaaS de GenAI, les fonctionnalités d'IA intégrées aux plateformes d'entreprise et les outils d'IA utilisés par les développeurs.

ARBITRAGES

Commencer par des alertes plutôt que par un blocage systématique se veut plus fluide et permet d'identifier les faux positifs des règles DLP avant qu'elles ne bloquent des activités légitimes.
En contrepartie, des données peuvent rester exposées pendant cette phase d'alerte. Tout dépend du niveau de risque que vous êtes prêt à accepter pour vos données sensibles ("red data"). Si une fuite de code source constitue un risque critique, il est préférable de bloquer dès le départ tout code détecté, quitte à accepter davantage de faux positifs.

Livrables à 30 jours

- Inventaire des applications d'IA, classées par niveau de risque (autorisées, tolérées ou interdites)
- Premières politiques d'accès en opération pour les applications d'IA et les groupes d'utilisateurs les plus à risque
- DLP activée pour les catégories de données sensibles et optimisation
- Premier tableau de bord décisionnel présentant le périmètre de l'IA et les indicateurs de risque
- Publication du Guide d'utilisation de l'IA et de la procédure de demande de dérogation

Jours 31 à 60

Rendre les politiques plus matures

Les jours 31 à 60 s'appuient sur la visibilité et les premières mesures de contrôle mis en place afin d'établir un modèle de politiques mature, des processus de gouvernance opérationnels et un inventaire de l'infrastructure d'IA.

Checklist des 60 premiers jours

Des politiques d’accès plus matures

- **Définir des politiques d’accès à l’IA propres à chaque département métier :** les RH, la finance, l’ingénierie et les ventes ont des besoins différents en matière d’outils d’IA et n’exposent pas les mêmes types de données
- **Ajouter des contrôles d’accès conditionnels :** niveaux d’accès distincts selon que le terminal d'accès est géré ou personnel (BYOD), critères géographiques d’accès et évaluation du risque liée à la posture de sécurité des terminaux
- **Formaliser trois niveaux d’accès pour les applications d’IA :** autorisées (accès complet avec inspection inline), tolérées (accès surveillé avec utilisation restreinte) et interdites (blocage avec possibilité de dérogation)
- Réexaminer les demandes de dérogation formulées au cours des 30 premiers jours et supprimer celles qui ne se justifient plus

Mettre en place des processus de gouvernance

- **Aligner les exigences de gouvernance de l’IA sur les politiques internes et les cadres réglementaires en vigueur :** NIST AI RMF, EU AI Act, référentiels de la CNIL pour les données de santé
- **Formaliser le processus de validation des outils d’IA :** soumission, examen de sécurité, classification des risques, validation conditionnelle et revalidation périodique
- Normaliser les niveaux de confidentialité pour les types de données utilisés par les applications d’IA, ainsi que des règles de traitement claires pour chaque type
- **Définir des procédures d’escalade pour les incidents de sécurité liés à l’IA :** désigner les responsables de la réponse aux incidents, préciser les incidents nécessitant une notification à la direction générale et définir les éléments de preuve à conserver

Tenir l'inventaire des ressources de l’infrastructure d’IA

- Inventorier les modèles, agents, services, jeux de données, bases vectorielles et pipelines d’IA, puis établir une nomenclature des composants d’IA (AI-BOM) en assurant la traçabilité depuis les jeux de données jusqu’à leur utilisation en environnement de production.
- **Évaluer la gestion de la posture de sécurité de l'IA (AI-SPM pour AI Security Posture Management) :** identifier les erreurs de configuration, les autorisations excessives, les données sensibles exposées et les vulnérabilités affectant les composantes de l’infrastructure d’IA
- **Traiter en priorité les problématiques AI-SPM les plus critiques :** autorisations excessives, données sensibles exposées et chemins d’accès non contrôlés entre les composants d’IA.
- **Vérifier que les principales plateformes d’IA sont prises en compte :** modèles hébergés dans le cloud, plateformes d’agents déployées en entreprise, intégrations de serveurs MCP et infrastructures d’IA non gérées

Protéger les applications d’IA privées

- **Ajouter la détection des injections de prompts aux mesures de contrôle pour les applications d’IA privées :** entrées malveillantes pour contourner les instructions du modèle ou exfiltrer des informations sensibles
- **Surveiller tout empoisonnement de données dans les pipelines RAG :** détecter les modifications non autorisées du corpus documentaire susceptibles d’altérer les réponses du modèle
- **Activer la journalisation des prompts et des réponses des systèmes d’IA privés avec détection des anomalies :** schémas de requêtes inhabituels, contenus inattendus dans les réponses et interactions en dehors du périmètre prévu
- Mettre en place des workflows d’examen des interactions avec les modèles pour les applications d’IA privées les plus sensibles

Livrables à 60 jours

- Modèle mature de politiques d'accès fondées sur les rôles, avec mise en oeuvre des accès conditionnels
- Processus de gouvernance de l'IA opérationnels : validation des outils, classification des données par niveau de confidentialité et processus d'escalade définis
- Visibilité établie sur l'AI-SPM et plan de remédiation priorisé disponible
- AI-BOM répertoriant les principaux composants de l'infrastructure d'IA
- Protections des applications d'IA privées déployées pour les applications internes les plus prioritaires
- Cadence du reporting de conformité définie et premier rapport livré

Jours 61 à 90

Automatiser et valider

Les jours 61 à 90 visent à rendre le programme autonome : application automatisée des politiques, surveillance continue de la conformité et validation par test de red teaming.

Checklist des 90 premiers jours

Automatiser l'application des politiques

- **Automatiser l'accompagnement des utilisateurs en cas de violation des politiques :** afficher un avertissement accompagné de recommandations adaptées au contexte, en amont de tout blocage
- **Intégrer la création de tickets ITSM ou SOAR pour les violations répétées :** créer automatiquement un ticket d'investigation lorsqu'un utilisateur déclenche à plusieurs reprises les mêmes alertes de contrôle
- **Configurer des mécanismes automatiques de quarantaine et d'isolation pour les comportements à risque :** soumission de volumes importants de données sensibles, accès à des destinations IA interdites et prompts suspects
- **Mettre en place un feedback :** suivre l'évolution des violations de politiques afin d'identifier les comportements qui incitent à ajuster les politiques plutôt que de renforcer les contrôles

Assurer un reporting de conformité fiable

- **Configurer une surveillance continue de l'état de conformité de l'IA :** identifier les contrôles actifs, les ressources prises en compte et les éventuelles lacunes dans la couverture
- **Assurer un mapping entre les mesures de contrôle et les cadres réglementaires en vigueur :** NIST AI RMF (y compris NIST AI 600-1 pour les risques liés à l'IA générative), EU AI Act, RGPD ou encore référentiels de la CNIL
- **Créer des packages de reporting prêts pour des audits :** synthèse de l'utilisation des applications d'IA, actions et résultats de la DLP appliquée aux prompts, dérogations accordées ou arrivées à expiration et état d'avancement de la remédiation des problématiques d'AI-SPM
- **Instaurer une revue trimestrielle du programme de sécurité de l'IA :** performances des politiques, lacunes de couverture, nouveaux domaines de risque et ajustements de la feuille de route

Red teaming et durcissement

- **Mettre en place un programme de red teaming pour les applications d'IA développées en interne :** bibliothèques de scénarios d'attaque prédéfinis complétées par des cas de test adaptés au contexte de chaque application
- **Tester l'ensemble de la surface d'attaque de l'IA :** injection de prompts, jailbreaks, empoisonnement du contexte, exfiltration de données personnelles, fuite de code source, comportements hors sujet et performances du modèle par rapport à des valeurs de référence
- **Déployer des politiques de protection pour les systèmes d'IA en production :** détection des injections de prompts, des fuites de données (personnelles et de code source), des tentatives de jailbreak et des violations des règles de modération des contenus
- **Automatiser la conversion des résultats du red teaming en des garde-fous dans l'environnement de production :** la différence entre les résultats des tests et l'application des contrôles en environnement de production révèle des points faibles dans la plupart des programmes de sécurité

16 min	72 %	100 %
Délai médian avant la première défaillance critique lors des tests de red teaming sur l'IA. Plus de 222 000 attaques menées sur plus de 25 environnements	des entreprises présentaient une vulnérabilité critique dès le premier test. Un risque critique est présent dès le premier jour, même dans les environnements matures	des systèmes d'IA testés présentaient au moins une vulnérabilité critique. Aucun système d'IA n'est sûr par défaut. Des tests continus sont indispensables.

Valider la conformité de l’architecture

- **Vérifier que tous les chemins d’accès à l’IA sont conformes aux six principes du Zero Trust :** vérifier que chaque principe est appliqué aux accès des utilisateurs, des développeurs et de l’infrastructure d’IA
- **Identifier et pallier les lacunes des politiques aux intersections** , à savoir les zones où une couche de contrôle s’arrête et où une autre prend le relais, avec des périmètres non couverts à l'intersection
- Normaliser l’application des politiques au moyen d’un modèle de politique unifié qui porte sur la vérification des identités, le contrôle des destinations, l’évaluation des risques, les mesures de contrôle inline, la réduction des privilèges d'accès et de l’exposition, ainsi que la surveillance

Livrables à 90 jours

- Réponse automatisée aux violations et coaching des utilisateurs en place
- Surveillance continue de la conformité opérationnelle, avec conformité documentée aux frameworks
- Reporting de gouvernance destiné aux décideurs et aux audits
- Programme de red teaming établi, avec politiques de protection actives en environnement de production
- Conversion automatisée des résultats du red teaming en garde-fous mis en oeuvre
- Scorecard du programme de sécurité et feuille de route trimestrielle pour une gouvernance en continu

Comment Zscaler sécurise l'adoption de l'IA

La plateforme Zero Trust Exchange™ est conçue pour la gouvernance et la sécurisation de l'IA à l'échelle de l'entreprise. Elle capitalise sur la plateforme qui gère déjà les accès des utilisateurs, des instances et au réseau pour des milliers d'entreprises dans le monde. Elle ne donne donc pas lieu à une rétro-ingénierie d'une architecture de sécurité déjà existante.

Chaque étape du cycle de sécurité de l'IA correspond à un ensemble spécifique de fonctionnalités de la plateforme. Le tableau ci-dessous présente les fonctionnalités offertes par Zscaler pour chaque étape.

IDENTIFIER Identifier l'intégralité de votre empreinte IA	**CONTRÔLER** Appliquer l'accès Zero Trust pour l'IA	**PROTÉGER** Durcir et protéger les applications d'IA
Gestion des ressources d'IA Zscaler identifie l'ensemble des applications, modèles, agents, serveurs MCP et fonctionnalités d'IA dans un SaaS, y compris les outils que les équipes IT n'ont jamais validé.	**Sécurité des accès pour l'IA** Appliquer des contrôles d'accès Zero Trust aux services SaaS avec IA, aux fonctionnalités d'IA intégrées aux plateformes d'entreprise et aux environnements de développement avec IA, avec inspection inline à chaque niveau.	**Red teaming et garde-fous pour l'IA** Tester les applications d'IA développées en interne dans des conditions reproduisant des attaques réelles, puis convertir automatiquement les résultats en politiques de protection en environnement de production.
• Identification de l'IA fantôme et visibilité sur son utilisation auprès de l'ensemble des utilisateurs • **Nomenclature des composants d'IA (AI-BOM) :** traçabilité depuis les jeux de données jusqu'aux modèles, aux agents et à leur utilisation en production • **AI-SPM :** évaluation continue de la posture de l'IA à la recherche d'erreurs de configuration, d'autorisations excessives et de données sensibles exposées • Détection des agents d'IA présents dans les applications déployées en entreprise ou fournies en mode SaaS	• Politiques granulaires d'autorisation, d'avertissement, d'isolation et de blocage selon le groupe d'utilisateurs, l'application et le type de données • DLP inline appliquée aux prompts, aux réponses et aux téléversements de fichiers sur l'ensemble du trafic IA • Navigateur Zero Trust Browser pour l'accès aux outils d'IA à partir de dispositifs non gérés ou BYOD • Extraction et classification des prompts pour les principales applications de GenAI	• **Red teaming automatisé et continu :** injection de prompts, jailbreaks, empoisonnement du contexte, fuites de données et comportements inhabituels • Durcissement des prompts et évaluation des modèles, avec mapping de la gouvernance avec les frameworks NIST AI RMF et EU AI Act, et avec le Top 10 OWASP des risques pour les LLM • **Garde-fous en production :** détection des jailbreaks, des fuites de données personnelles et de code source, des violations des règles de modération de contenus et des comportements hors sujet • **Génération de politiques en boucle fermée :** les résultats des tests de red teaming sont automatiquement transformés en capteurs déployés en environnement de production

L'AVANTAGE DE LA PLATEFORME

La plupart des fournisseurs de solutions de sécurité ne couvrent qu'un volet de la problématique de la sécurité de l'IA. Zscaler protège l'ensemble du cycle de vie de l'IA à partir d'une seule plateforme : identification de toutes les applications d'IA, contrôle des accès par Zero Trust ainsi que la protection des applications et de l'infrastructure d'IA du développement jusqu'à la mise en production. La boucle fermée qui relie les résultats du red teaming aux garde-fous actifs en production est une capacité qu'aucune solution autonome ne peut reproduire.

Foire aux questions

Que signifie concrètement le « Zero Trust IA » ?
Le Zero Trust IA consiste à appliquer le principe fondamental du Zero Trust (absence de confiance implicite, validation continue) à chaque interaction avec l'IA au sein de votre environnement. Concrètement, aucune application d'IA, aucun utilisateur, aucun terminal, aucun prompt et aucun connecteur n'est considéré comme étant de confiance par défaut. L'accès est accordé en fonction de l'identité vérifiée, de l'état de sécurité du terminal, de la sensibilité des données et du contexte de la session. Les mesures de contrôle sont appliquées inline et en temps réel. Chaque décision d'accès est journalisée et traçable.

Que faire au cours des 30 premiers jours pour réduire les risques liés à l'IA fantôme ?
Commencez par l'identification. Vous ne pouvez gérer une IA que vous ne voyez pas. ThreatLabz recense plus de 3 400 applications d'IA dans les flux des entreprises, un chiffre qui a quadruplé en un an, principalement sous l'effet de fonctionnalités d'IA intégrées aux outils déjà approuvés par les équipes IT. Une fois cette visibilité acquise, répertoriez les applications les plus risquées et définissez des politiques de destination. Activez ensuite la DLP pour les prompts utilisant vos données confidentielles les plus à risque. Enfin, publiez un guide d'utilisation clair ainsi qu'une procédure de demande de dérogation afin d'offrir aux utilisateurs une solution conforme, plutôt que de les inciter à contourner les dispositifs de contrôle en place.

Comment les contrôles d'accès à l'IA et la DLP inline protègent-ils les prompts et les téléversements ?
Les contrôles d'accès déterminent quels utilisateurs peuvent accéder à quels outils d'IA, à partir de quels terminaux, depuis quels emplacements et dans quelles conditions. La DLP inline intervient sur le contenu en inspectant, en temps réel, les données effectivement soumises par les utilisateurs : texte des prompts, téléversements de fichiers et opérations de copier-coller. Ces deux mécanismes sont complémentaires : les contrôles d'accès déterminent qui peut accéder à quels outils d'IA, tandis que la DLP contrôle les données qui transitent par les canaux autorisés. Une vulnérabilité dans l'un ou l'autre expose les données à des risques. C'est pourquoi les entreprises qui n'appliquent la DLP qu'aux téléversements de fichiers, sans inspecter les prompts, restent exposées aux fuites de données.

Qu'est-ce que l'AI-SPM et pourquoi est-il essentiel à l'adoption de l'IA ?
L'AI-SPM, à savoir la gestion de la posture de sécurité de l'IA, consiste à évaluer en continu les configurations, les autorisations et les risques d'exposition au sein de l'infrastructure d'IA. Elle constitue l'équivalent, pour l'IA, du CSPM appliqué aux infrastructures cloud. Elle recense les ressources d'IA, évalue leurs configurations au regard des bonnes pratiques de sécurité, identifie les autorisations excessives et l'exposition de données sensibles, puis suit l'évolution de la posture de sécurité dans le temps. Cette approche est essentielle, car les erreurs de configuration dans les infrastructures d'IA peuvent être nombreuses et mal maîtrisées par la plupart des entreprises qui passent de l'expérimentation de l'IA à son déploiement.

Quels sont les indicateurs clés de performances (KPI) qui démontrent le mieux à une direction générale que les risques liés à l'IA diminuent, mais sans nuire à la productivité ?
Les KPI les plus pertinents sont un indicateur de réduction des risques et un indicateur d'adoption qui permettent à des dirigeants de suivre simultanément ces deux tendances. La diminution du nombre d'applications d'IA fantôme parallèlement à une adoption croissante des applications d'IA autorisées constitue un signe d'une gouvernance efficace. Les actions de DLP appliquées aux prompts (blocage, avertissement ou autorisation) indiquent si les mesures de contrôle sont correctement calibrées. Le délai de traitement des demandes de dérogation et le délai de validation des outils d'IA reflètent quant à eux la réactivité de la gouvernance.

Act Fast. Stay Secure.

À propos de Zscaler

Zscaler (NASDAQ : ZS) est un pionnier et un leader mondial de la sécurité par Zéro Trust. Les plus grandes entreprises mondiales, les opérateurs d'infrastructures critiques et les administrations comptent sur Zscaler pour sécuriser leurs utilisateurs, leurs sites distants, leurs applications, leurs données et leurs appareils, mais aussi pour accélérer leurs initiatives de transformation digitale. Adossée à plus de 160 data centers dans le monde, la plateforme Zscaler Zero Trust Exchange™, optimisée par une IA avancée, déjoue chaque jour des milliards de cybermenaces et violations de politiques de sécurité et assure des gains de productivité pour les entreprises modernes en réduisant les coûts et la complexité.

+1 408 533 0288 **Zscaler, Inc. (siège) • 120 Holger Way • San Jose, CA 95134** **zscaler.com/fr**