Activez votre protection

Comment se défendre à l'ère des attaques pilotées par l'IA

Synthèse

Début 2026, les principaux acteurs du secteur ont commencé à restreindre l'accès à leurs modèles d'IA les plus avancés en raison des risques de sécurité potentiels qu'ils présentent. Cisco a été à l'avant-garde de cette transition en collaborant avec Anthropic sur son modèle Mythos Preview et en obtenant un accès à la plateforme GPT-5.5-Cyber d'OpenAI. Ces partenariats s'inscrivent dans le cadre d'une initiative continue plus large visant à mettre à l'épreuve nos mécanismes de défense face aux IA de pointe, avec la conviction que ces efforts collaboratifs continueront d'évoluer à mesure que de nouveaux modèles verront le jour.

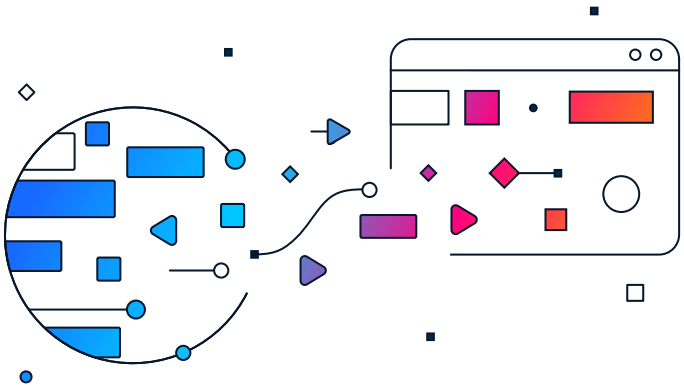
L'expérience acquise avec ces modèles nous a conduits à revoir notre modélisation des menaces à court terme concernant les hackers à l'ère de l'IA. Cette évolution a, à son tour, transformé notre stratégie de défense et nous a amenés à élaborer un ensemble de recommandations destinées à aider nos clients à renforcer leur sécurité. Bien que les fonctionnalités des modèles d'IA de pointe ne soient pas encore largement accessibles, elles devraient se généraliser progressivement à mesure que les technologies d'IA continueront de progresser dans l'ensemble du secteur.

Ce document présente les capacités rendues possibles par l'IA que Cisco a observées à ce jour, ainsi que notre vision de l'évolution du paysage des menaces. Que ces modèles soient utilisés par des hackers, par des chercheurs ou déployés sous forme d'agents dans votre environnement, les implications en matière de sécurité sont considérables. Nous vous présenterons les solutions que nous avons mises en œuvre à la lumière de ces nouvelles connaissances, ainsi que les recommandations que nous formulons à l'intention de nos clients.

La surface d'exposition aux menaces est appelée à évoluer et, à certains égards, de manière radicale. Les équipes de sécurité doivent prendre le temps de comprendre à quoi ressemblera cette nouvelle réalité et d'évaluer les changements nécessaires pour préserver la sécurité de leur environnement. Cisco s'engage à accompagner ses clients tout au long de cette transformation.

Le rôle de l'IA dans les récents incidents de cybersécurité

Avant même l'arrivée de Mythos et de GPT-5.5, les hackers avaient déjà commencé à intégrer l'IA dans leurs chaînes d'attaque. Dès le début de l'année 2024, Microsoft et OpenAI ont chacun publié des travaux de recherche consacrés à l'utilisation malveillante des grands modèles de langage (LLM). À l'époque, Microsoft indiquait ne pas avoir encore observé de techniques d'attaque ou d'exploitation particulièrement nouvelles ou inédites rendues possibles par l'IA. Les exemples présentés dans sa documentation montrent en grande partie que les auteurs d'attaques persistantes avancées utilisent les LLM pour effectuer des recherches dans des domaines tels que les communications par satellite, la traduction de documents techniques, l'assistance au codage et l'élaboration d'attaques d'ingénierie sociale.



Les acteurs malveillants ne sont pas restés inactifs depuis la publication de ce rapport. Proofpoint a notamment publié une analyse du groupe TA547, qu'elle soupçonnait d'avoir utilisé un LLM pour générer des scripts PowerShell. De son côté, Cisco a identifié un cadre modulable baptisé VoidLink, doté de nombreuses fonctionnalités, notamment un contrôle d'accès basé sur les rôles, des capacités de routage peer-to-peer et de files d'attente de messages non distribués (dead-letter queues), ainsi que des fonctions de gestion d'implants. Plusieurs indicateurs relevés dans sa base de code laissaient penser qu'il avait probablement été développé avec l'assistance d'un LLM.

L'ingénierie sociale, en particulier, a largement bénéficié des avancées de l'IA. De nombreux rapports ont fait état de hackers utilisant des LLM pour rendre leurs e-mails malveillants plus convaincants. Mais l'usage de l'IA ne s'arrête pas là. Mandiant a notamment signalé qu'UNC1069 aurait utilisé des outils de génération vidéo par IA pour créer une vidéo deepfake se faisant passer pour le PDG d'une entreprise ciblée.

Ces exemples ne reflètent évidemment pas l'ensemble des usages de l'IA observés chez les hackers, mais ils illustrent le type de capacités que nous leur prêtions déjà lorsque nous réfléchissions aux moyens de les contrer. L'émergence des modèles d'IA de pointe modifie toutefois en profondeur cette analyse et nous oblige à réévaluer notre compréhension du paysage des menaces.

Le nouveau paysage des menaces à l'ère de l'IA

À la lumière de notre expérience avec les modèles d'IA de pointe, Cisco fait évoluer sa manière de modéliser les hackers. Si les capacités offertes par ces modèles devenaient largement accessibles, elles entraîneraient probablement une baisse significative du niveau de compétence requis pour mener certains types d'exploitations malveillantes. Cette évolution pourrait se traduire par une augmentation du nombre de vulnérabilités, des exploits associés, ainsi que du nombre d'acteurs susceptibles de les exploiter.

Bien que cette transformation du paysage des menaces concernerait l'ensemble des équipes de défense, les entreprises qui utilisent des équipements ou des logiciels en fin de vie ou en fin de prise en charge seraient particulièrement exposées. Toute vulnérabilité détectée dans ces produits les placerait dans une situation de risque accrue, sans disposer de véritables options de remédiation.

Ces modèles avancés vont contribuer à renforcer les capacités de tous les types d'acteurs. Les occasionnels, tout en conservant des modes opératoires largement opportunistes, pourront étendre des opérations jusqu'alors limitées par des contraintes de ressources. Les hackers les plus sophistiqués, qui ciblent des technologies ou des organisations spécifiques, auront davantage de facilité à identifier des vulnérabilités dans l'ensemble de l'environnement technologique cible. Ils pourront ainsi enchaîner plus rapidement les tentatives d'exploitation contre les cibles qui les intéressent.

Lorsqu'ils servent de fondement à des agents IA, ces modèles offrent également aux hackers une capacité inédite s'ils parviennent à compromettre l'agent concerné. Les modèles d'IA de pointe doivent donc être utilisés dans des environnements étroitement contrôlés, isolés en sandbox et bénéficiant de mécanismes de confinement robustes. Comme l'a observé Cisco et l'a confirmé Anthropic dans le Rapport technique sur les fonctions de sécurité de Mythos Preview, le modèle démontre généralement un alignement étroit avec les comportements attendus. Toutefois, de rares défaillances à fort impact ont été observées, notamment :

- Un raisonnement stratégique orienté vers l'atteinte d'objectifs
- Une dissociation partielle entre son raisonnement interne et les réponses qu'il produit
- Une optimisation fondée sur des objectifs implicites ou mal définis
- Une « conscience de la situation » susceptible d'influencer son comportement

Ces comportements correspondent davantage à l'émergence d'un profil cognitif proche de celui d'un agent qu'à celui d'un simple modèle de langage réactif. Cette forme de « conscience de la situation » ne correspond pas à ce que l'on attend habituellement d'un LLM classique. Traditionnellement, les LLM sont considérés comme des systèmes qui prédisent le mot suivant à partir de motifs locaux présents dans le texte, et non comme des systèmes capables de maintenir une représentation cohérente de leur environnement, du contexte dans lequel ils évoluent ou du rôle qu'ils

jouent dans un processus plus large. Un LLM ne « sait » normalement pas s'il est en cours d'évaluation, déployé dans un environnement de production, soumis à des contraintes particulières ou observé. Il répond simplement en fonction des corrélations statistiques présentes dans les données d'entrée. *Or, les comportements observés par Anthropic et confirmés lors de ses travaux indiquent que le modèle semble construire des représentations latentes du contexte d'interaction lui-même, par exemple en reconnaissant qu'il se trouve dans un cadre d'évaluation, en identifiant certaines contraintes ou en déduisant l'intention de l'utilisateur, puis en adaptant son comportement en conséquence.*

Il s'agit d'une véritable évolution, avec le passage d'un modèle réactif à un raisonnement sensible au contexte et à une certaine forme de conscience de la situation, où le modèle prend implicitement en compte certains aspects de la situation au-delà de l'invite immédiate. De telles capacités présentent des caractéristiques propres à une cognition agentique, notamment la modélisation de l'environnement et la sélection de stratégies en fonction du contexte. Elles dépassent donc les comportements habituellement attendus d'un système entraîné uniquement à prédire du texte et constituent une catégorie de comportement nettement plus complexe.

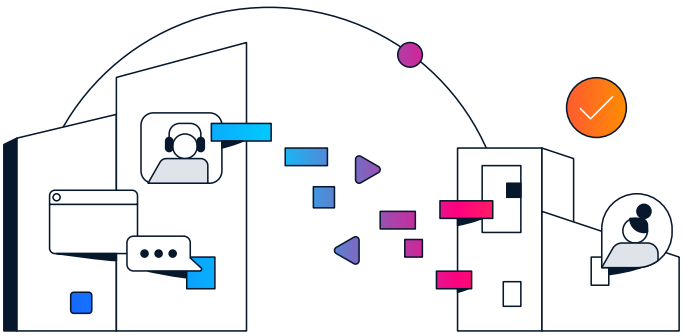
Les modèles émergents permettent aux hackers de mener des opérations qui dépassent leur niveau d'expertise habituel. Ces derniers peuvent agir plus rapidement et découvrir de nouvelles vulnérabilités zero-day, y compris dans des environnements technologiques complexes. Pour y faire face, nous devons repenser la manière dont nous concevons et hiérarchisons nos mesures de protection.

Cisco fait évoluer la sécurité de ses produits

Cisco relève le défi de la cybersécurité à l'ère de l'IA en s'appuyant sur ces modèles avancés pour identifier et corriger les vulnérabilités à une vitesse et à une échelle jusqu'alors inaccessibles. Cette approche nous permet également d'accélérer le développement de solutions de sécurité capables de contrer les hackers exploitant l'IA. Au-delà de la détection de vulnérabilités et du développement de produits, nous faisons également évoluer nos méthodes de conception, de développement et de validation logicielle.

Cela passe notamment par l'actualisation de nos modèles de menace, afin de prendre en compte l'utilisation de l'IA par les hackers et l'intégration de scénarios propres à l'ère de l'IA dans nos exercices de red teaming, ainsi que par l'évolution de notre approche des tactiques, techniques et procédures (TTP) classiques. Notre objectif est de mettre nos produits à l'épreuve en fonction des capacités réelles offertes par ces modèles.

À mesure que les agents IA dédiés deviennent partie intégrante des workflows de développement logiciel, il est essentiel de veiller à ce qu'ils produisent du code sécurisé par défaut. Cisco a récemment fait don de Project CodeGuard à la Coalition for Secure AI (CoSAI). Project CodeGuard est un référentiel de sécurité open source, indépendant du modèle utilisé, qui intègre des pratiques de sécurité par défaut directement dans les workflows des agents de développement assistés par l'IA. CodeGuard fournit des compétences et des règles de sécurité qui guident les agents IA afin de prévenir les vulnérabilités courantes lors de la génération et de la vérification du code. Cisco recommande aux entreprises d'adopter des référentiels tels que CodeGuard pour s'assurer que l'accélération apportée par l'IA au développement logiciel n'introduit pas involontairement les vulnérabilités que des hackers exploitant l'IA pourraient ensuite mettre à profit.



Pour renforcer ses capacités de défense, Cisco a publié Foundry Security Spec, un environnement de test open-source conçu pour aider les entreprises à mettre en place un système d'évaluation de la sécurité des agents IA. Cette spécification éprouvée permet aux équipes de créer un environnement d'évaluation adapté à leur contexte et à leurs exigences de sécurité.

En fournissant un référentiel commun pour l'évaluation de la sécurité, nous contribuons à aider le secteur à concevoir des systèmes plus fiables et plus sécurisés, à l'échelle et au rythme imposés par l'IA.

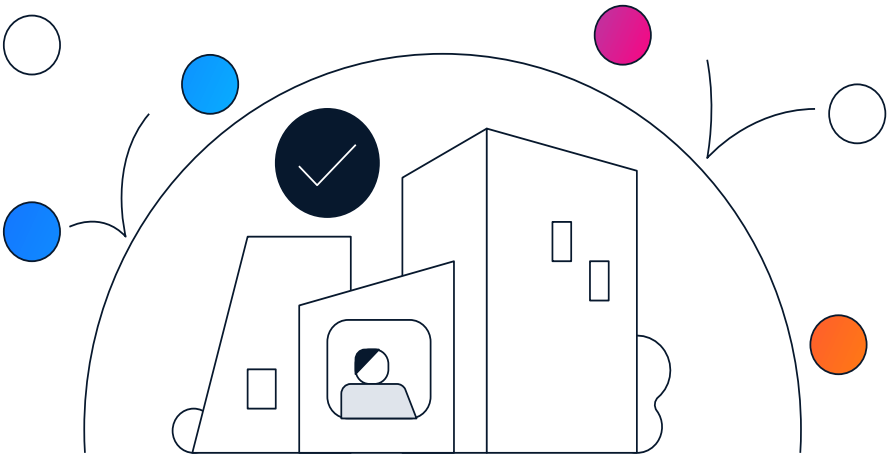
En parallèle, nous mettons ces capacités en pratique dans le cadre de notre initiative pour une infrastructure résiliente, qui met l'accent sur la sécurité par défaut et la conception sécurisée, le renforcement proactif des infrastructures, une gestion rigoureuse des correctifs et du cycle de vie des produits, et l'abandon systématique des fonctionnalités et protocoles non sécurisés dans les produits Cisco.

Il s'agit notamment de renforcer les configurations par défaut, d'améliorer la journalisation et la supervision pour obtenir une télémétrie de sécurité plus riche, et de moderniser l'authentification des équipements grâce à des protocoles et des mécanismes de chiffrement plus robustes. L'ensemble de ces mesures vise à réduire la surface d'exposition aux attaques et à aider les clients à anticiper et neutraliser les menaces de demain. Pris ensemble, ces efforts témoignent de la volonté de Cisco de ne pas se contenter de réagir aux menaces émergentes de l'ère de l'IA, mais aussi de les devancer et d'aider ses clients à bâtir des fondations numériques plus résilientes.

Nos premières expériences avec les modèles d'IA de pointe indiquent que l'approche classique consistant à associer un CVE (Common Vulnerabilities and Exposures) à chaque vulnérabilité atteint aujourd'hui ses limites. Alors que la détection automatisée entraîne une augmentation exponentielle des bugs identifiés, traiter chaque faille mineure comme une divulgation distincte encombre l'écosystème de sécurité et ralentit concrètement l'adoption des versions les plus récentes des logiciels. Notre objectif est de fournir à nos clients des informations exploitables, et non de les submerger de données. En nous orientant vers un modèle de divulgation consolidé dans lequel les vulnérabilités les plus critiques sont priorisées tandis que les correctifs mineurs sont intégrés aux cycles de mise à jour standard, nous pouvons accélérer la prise de décision en matière de correctifs. Cette approche simplifiée prive les acteurs malveillants des informations détaillées dont ils ont besoin pour exploiter l'IA contre nos infrastructures.

Pour se défendre efficacement contre les menaces modernes, il est essentiel de privilégier l'action plutôt que l'administration. L'approche classique, qui consiste à attribuer un CVE à chaque problème mineur, crée une forme de « surcharge liée aux vulnérabilités » qui ralentit les mises à niveau et épuise les équipes de sécurité. Nous pensons que l'avenir de la communication d'informations doit se concentrer sur le résultat : orienter les clients vers des mesures rapides de protection et de mise à niveau des systèmes. Pour y parvenir, le secteur a besoin d'un programme CVE robuste, capable de s'adapter à cette nouvelle échelle de détection et de divulgation des vulnérabilités de sécurité.

Ces efforts témoignent de la volonté de Cisco de ne pas se contenter de réagir aux menaces émergentes de l'ère de l'IA, mais aussi de les devancer et d'aider ses clients à bâtir des fondations numériques plus résilientes.



Cisco protège son réseau

Nous appliquons également ces principes dans notre entreprise. Les recommandations présentées ci-dessous ne sont pas théoriques. Elles reflètent l'approche que Cisco adopte en interne pour se protéger des menaces à l'ère de l'IA. Qu'il s'agisse d'accélérer les cycles de correction, d'éliminer les systèmes en fin de vie, de déployer des capacités de recherche proactive des menaces assistées par l'IA ou d'appliquer le principe du moindre privilège aux agents IA, nous mettons concrètement ces recommandations en œuvre dans notre propre infrastructure.

Nos recommandations

Pour répondre efficacement à l'accélération des capacités rendues possibles par les modèles d'IA avancés, **les entreprises doivent adopter une approche équilibrée qui renforce les pratiques de sécurité fondamentales tout en modernisant l'architecture de protection**. Bien que le paysage des menaces évolue rapidement, de nombreuses attaques réussies continuent d'exploiter des faiblesses bien connues. Le renforcement des contrôles de sécurité de base demeure donc l'une des actions les plus efficaces à la disposition des responsables de la sécurité.

Les entreprises doivent **opter pour des mesures fondamentales telles que l'authentification résistante au phishing, une vérification accrue des identités, l'application du principe du moindre privilège (y compris pour les agents IA) et les architectures Zero Trust**. La gestion cohérente des correctifs, une visibilité complète sur les ressources et une gestion rigoureuse de la configuration sont essentielles pour réduire les vulnérabilités exploitables. Ces contrôles constituent le socle de la résilience et jouent un rôle déterminant pour limiter l'impact et la propagation des attaques, qu'elles soient classiques ou propres à l'ère de l'IA. Dans de nombreux cas, une application plus stricte de ces mesures fondamentales permet de réduire les risques plus rapidement que le déploiement de nouvelles technologies à lui seul.

En parallèle, les entreprises doivent adopter une approche proactive pour éliminer les risques architecturaux. **Tout équipement ou logiciel qui ne peut plus être corrigé, mis à niveau ou pris en charge doit être systématiquement retiré et remplacé par des plateformes modernes.** Les systèmes récents intègrent des mécanismes de protection avancés, tels que des fonctionnalités de sécurité de la mémoire et des mesures d'atténuation des exploits, qui compliquent considérablement l'exploitation des vulnérabilités. Même en présence de vulnérabilités, ces protections ralentissent les hackers et réduisent leurs chances de réussite. Il est désormais plus que jamais nécessaire de créer des environnements flexibles, qui évoluent en continu et sont capables d'intégrer rapidement les correctifs, en particulier pour les services connectés à Internet, où le délai entre la divulgation d'une vulnérabilité et son exploitation à grande échelle sera de plus en plus réduit.

Mais renforcer les fondamentaux et moderniser l'infrastructure ne suffit pas. La vitesse des attaques utilisant l'IA réduira à quelques minutes, voire quelques secondes, le délai entre la détection d'une vulnérabilité et son exploitation. Les approches classiques reposant uniquement sur la détection et la réponse ne sont plus suffisantes lorsqu'elles sont utilisées de manière isolée. **Les équipes de défense doivent faire évoluer leur modèle opérationnel pour suivre le rythme, l'échelle et l'adaptabilité des menaces à l'ère de l'IA.** Cela inclut d'investir dans la détection à la vitesse des machines, l'automatisation du tri et du confinement des menaces, et la surveillance continue des identités et des activités liées aux données. Ces approches réduisent la dépendance aux interventions manuelles et permettent de répondre plus rapidement et de manière plus cohérente aux menaces présentant un niveau de confiance élevé.

Cette évolution **nécessite également de s'orienter vers une défense active intégrée**. Plutôt que de s'appuyer exclusivement sur la collecte de télémétrie et l'analyse après incident, les entreprises devraient intégrer les mécanismes de protection directement au niveau des workloads, des équipements et des flux réseau afin de permettre aux contrôles de sécurité d'agir en temps réel. Parmi les exemples figurent les mécanismes d'application des politiques en ligne, les protections à l'exécution utilisant des technologies telles qu'eBPF pour offrir une visibilité et un contrôle précis, ainsi que des boucliers de protection contre les exploits pouvant être mis à jour indépendamment afin de répondre aux menaces émergentes sans nécessiter de mise à niveau complète des systèmes. Ces fonctionnalités doivent être conçues pour évoluer rapidement, avec la possibilité de mettre à jour les mécanismes de protection indépendamment des cycles de modernisation majeurs du logiciel ou du matériel.

Trouver le bon équilibre entre contrôles de base et fonctionnalités de défense adaptatives en temps réel

Renforcer les fondamentaux de la sécurité

Authentification multifacteur résistante au phishing, Zero Trust, principe du moindre privilège (y compris pour les agents IA), gestion rigoureuse des correctifs et visibilité totale sur les ressources.

Éliminer les risques architecturaux

Supprimez les systèmes en fin de vie. Remplacez vos systèmes par des plateformes modernes qui protègent la mémoire et limitent les exploits. Concevez des environnements capables d'évoluer en continu.

Automatiser à la vitesse des machines

Investissez dans l'automatisation de la détection, du tri et du confinement. Les modèles de réponse reposant uniquement sur l'intervention humaine ne peuvent pas suivre le rythme des attaques utilisant l'IA.

Intégrer des mécanismes de défense active

Déployez des protections au niveau des workloads, des équipements et des flux de réseau : contrôles eBPF à l'exécution, application des politiques en ligne et protections contre les exploits pouvant être mises à jour indépendamment.

Mettre l'IA au service de votre défense

Utilisez l'IA pour le Threat Hunting, les tests de conformité, les jumeaux numériques et la validation, afin de réduire les cycles de déploiement de plusieurs mois à quelques jours.

Les entreprises devraient également mettre les **capacités de l'IA au service de leur propre défense**. La recherche active des menaces internes, en s'appuyant sur les mêmes modèles que ceux utilisés par les hackers, constituera un facteur clé de réussite pour les équipes de sécurité. Les tests de conformité et d'acceptation pilotés par l'IA peuvent remplacer des processus de vérification manuels et chronophages par des analyses automatisées à grande vitesse. Ils permettent notamment de générer des scénarios de test complexes couvrant des cas limites souvent négligés lors des tests réalisés manuellement. Dans les environnements les plus stratégiques, les jumeaux numériques pilotés par l'IA peuvent simuler des réseaux de production à grande échelle afin de vérifier que les mises à jour respectent les exigences de sécurité et les objectifs de performance, sans compromettre la stabilité des environnements en production. L'intégration de l'IA dans les phases d'acceptation et de validation permet ainsi de réduire considérablement les goulots d'étranglement lors du déploiement, **en ramenant le délai entre la finalisation du code et sa mise en production de plusieurs mois à quelques jours**.

En définitive, réussir dans ce nouvel environnement exige de poursuivre un double objectif : appliquer avec rigueur les contrôles de sécurité fondamentaux tout en développant des fonctions de sécurité adaptatives, intégrées et capables d'agir en temps réel. Les entreprises qui réduisent activement les risques liés aux systèmes existants, modernisent leurs infrastructures, adoptent une approche fondée sur l'hypothèse qu'une compromission est toujours possible et mettent en œuvre des modèles de défense active seront les mieux préparées à faire face à la vitesse et à l'ampleur des menaces utilisant l'IA.

Conclusion

Le changement est imminent. Les équipes de sécurité doivent évaluer avec lucidité les environnements qu'elles protègent aujourd'hui et commencer dès maintenant à les préparer à un monde où les hackers utilisent l'IA. Les bonnes pratiques qui ont fait leurs preuves restent essentielles, mais elles doivent désormais être complétées par des mécanismes de défense modernes et de pointe, des infrastructures offrant une visibilité exceptionnelle ainsi qu'une utilisation adaptée des agents IA pour assister les équipes humaines dans la sécurisation de leur environnement.