Se protéger

Conseils pour se défendre à l'ère des attaques activées par l'intelligence artificielle

Synopsis

Au début de l'année 2026, les chefs de file du secteur ont commencé à restreindre l'accès à leurs modèles d'intelligence artificielle les plus avancés en raison des risques en matière de sécurité. Cisco a été à l'avant-garde de cette réorientation, en collaborant avec Antropic sur son modèle Mythos Preview et en obtenant un accès à GPT-5.5-Cyber d'OpenAI. Ces partenariats s'inscrivent dans le cadre d'une initiative continue plus vaste visant à mettre à l'essai nos défenses face à l'intelligence artificielle exploratoire, en reconnaissant que nos efforts de collaboration continueront d'évoluer à mesure que de nouveaux modèles voient le jour.

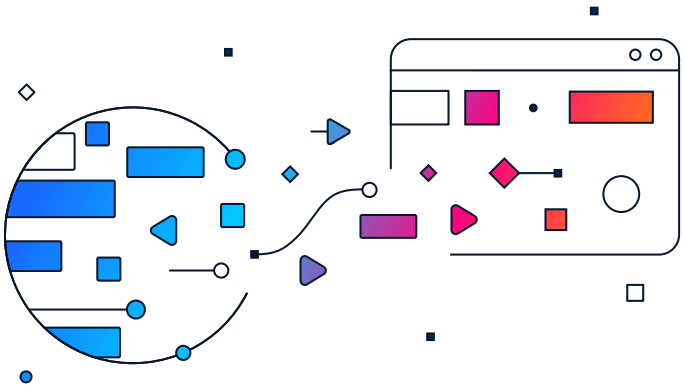
Notre expérience avec ces modèles nous a amenés à modifier notre modélisation des menaces à court terme liées aux adversaires de l'ère de l'intelligence artificielle. Cela a donc modifié notre manière de nous défendre et nous a amenés à élaborer un ensemble de recommandations préventives à l'intention des clients. Bien que les capacités des modèles d'intelligence artificielle exploratoire ne soient pas accessibles à grande échelle, nous prévoyons que ces capacités, et plus encore, le seront à mesure que cette technologie progresse à tous les niveaux.

Ce document présente les conclusions tirées par Cisco jusqu'à présent en ce qui concerne les fonctionnalités basées sur l'intelligence artificielle et ce à quoi, selon nous, ressemblera le nouveau cadre des menaces. Que ces modèles soient exploités par des agresseurs, utilisés par des chercheurs ou déployés en tant qu'agents dans votre propre environnement, les répercussions sur la sécurité sont importantes. Nous vous ferons part de ce que nous avons mis en œuvre en fonction de cette nouvelle compréhension et présenterons nos recommandations aux clients.

La surface des menaces va évoluer, et à certains égards, de façon spectaculaire. Les défenseurs doivent prendre le temps de comprendre à quoi ressemblera la nouvelle normalité et évaluer les modifications qu'ils doivent apporter à leur environnement pour rester sécurisés. Cisco s'engage à être un partenaire tout au long de cette transformation.

L'intelligence artificielle au cœur des événements récents en matière de cybersécurité

Avant même Mythos et GPT-5.5, les acteurs malveillants ont intégré l'intelligence artificielle à leurs flux d'attaques. Au début de l'année 2024, Microsoft et OpenAI ont chacun publié des recherches sur l'utilisation malveillante des grands modèles de langage (GML). À l'époque, Microsoft a déclaré « ne pas avoir encore observé de techniques d'attaque ou d'utilisation abusive particulièrement nouvelles ou uniques optimisées par l'intelligence artificielle ». Leur documentation montre en grande partie des acteurs de menaces persistantes avancées (MPA) qui utilisent les grands modèles de langage pour effectuer des recherches dans des domaines comme les communications par satellite, la traduction de documents techniques, l'assistance au codage et l'élaboration d'attaques d'ingénierie sociale.



Les acteurs ne sont pas demeurés inactifs depuis la publication de ce rapport. Proofpoint a publié un document sur TA547, où ils soupçonnaient cet acteur d'utiliser un GML pour générer des scripts PowerShell. De même, Cisco a identifié un cadre modulaire appelé VoidLink, un outil qui offre des capacités étendues comme le contrôle des accès basé sur les rôles, des capacités de routage entre pairs et de type files d'attente des messages non remis, ainsi que la capacité de gestion des implants. Un certain nombre d'indicateurs ont été trouvés dans la base de code, ce qui suggère qu'elle a probablement été développée avec l'aide d'un GML.

L'ingénierie sociale, en particulier, a profité de l'utilisation de l'intelligence artificielle. De nombreux rapports montrant des acteurs utilisant les GML pour améliorer les leurres par courriel ont été rédigés. Cependant, les acteurs sont allés bien au-delà, avec les rapports Mandiant sur l'utilisation potentielle par UNC1069 des outils vidéo basés sur l'intelligence artificielle pour créer une vidéo hypertruquée censée provenir du président-directeur général de l'entreprise cible.

Cela ne représente certainement pas toute l'utilisation de l'intelligence artificielle par les adversaires qui ont été observés, mais cela est illustratif du type de capacités que nous supposons que les acteurs avaient lorsque nous avons discuté de la manière de contrer les acteurs utilisant l'intelligence artificielle. Les capacités qu'apportent les modèles exploratoires changent nécessairement la façon dont nous évaluons le cadre des menaces.

Le nouveau cadre des menaces alimentées par l'intelligence artificielle

Fort de notre expérience de travail avec les modèles d'intelligence artificielle exploratoire, Cisco est en train de changer la manière dont nous modélisons nos adversaires. Si elles étaient largement accessibles, les fonctionnalités de ces modèles entraîneraient probablement une baisse drastique des compétences pour certains types d'activités d'exploitation. Cela entraînerait un plus grand nombre de vulnérabilités et d'exploits connexes ainsi qu'un ensemble plus large d'acteurs susceptibles de tirer parti de ces vulnérabilités.

Bien que la portée potentielle de ce changement de cadre concerne tous les défenseurs, ceux qui utilisent des appareils ou des logiciels en fin de vie ou en fin de soutien seraient particulièrement vulnérables. Les vulnérabilités découvertes dans ces produits rendraient les défenseurs particulièrement vulnérables et sans bonnes options de correctifs.

Ces modèles avancés offriront un renforcement des capacités pour tous les niveaux d'acteurs. Les acteurs du secteur des produits de base, tout en restant grandement opportunistes, auront la possibilité de faire évoluer des opérations qui étaient auparavant limitées en matière de ressources. Les acteurs de niveau supérieur avec un ciblage plus précis auront plus de facilité à découvrir les vulnérabilités de la pile technologique cible. Cela se traduira par une réduction du temps d'arrêt entre les tentatives d'exploit sur les cibles de préférence.

Ces modèles, lorsqu'ils sont utilisés comme base pour les agents d'intelligence artificielle, représentent une nouvelle capacité pour les agresseurs si ces derniers peuvent compromettre cet agent. Les modèles d'intelligence artificielle exploratrice doivent être exploités dans des environnements étroitement contrôlés et en bac de sable avec un confinement solide. Comme Cisco l'a observé et Anthropic l'a confirmé dans le rapport technique sur les capacités de sécurité de Mythos Preview, le modèle démontre une performance élevée d'alignement de référence, mais présente des défaillances rares et de gravité élevée classées par :

- Le raisonnement stratégique orienté sur les objectifs
- Le découplage partiel entre la cognition interne et les résultats
- L'optimisation vers des objectifs implicites ou mal précisés
- La « connaissance de la situation » qui influence le comportement

Ces comportements sont conformes à un profil cognitif émergent semblable à celui d'un agent, plutôt qu'à un modèle de langage purement réactif. Ce comportement de « connaissance de la situation » n'est pas ce à quoi nous nous attendons généralement d'un GML standard. Les grands modèles de langage traditionnels sont considérés comme des prédicteurs du jeton suivant opérant sur des modèles locaux dans le texte, et non comme des systèmes qui maintiennent un modèle cohérent de leur environnement, de leur contexte ou de leur rôle dans un processus

plus large. Un GML ne « sait » pas s'il est évalué, déployé, limité ou observé. Il répond simplement en fonction des corrélations statistiques dans l'entrée. *Mais les comportements observés et confirmés par Anthropic impliquent que le modèle forme des représentations latentes du contexte d'interaction lui-même (par exemple, en reconnaissant les paramètres d'évaluation, les contraintes ou l'intention de l'utilisateur) et ajuste son comportement en conséquence.*

Il s'agit certainement d'un virage d'une exécution de modèle purement réactive à un raisonnement contextuel et autoréférentiel, dans lequel le modèle suit implicitement les aspects de la situation au-delà de l'invite immédiate. Ces capacités ressemblent à des éléments de la cognition agentive (y compris la modélisation de l'environnement et la sélection de stratégies conditionnelles), qui vont au-delà du comportement attendu d'un système formé uniquement pour prédire le texte et représentent donc une classe de comportements de modèle qualitativement différente et plus complexe.

Les modèles émergents permettent aux agresseurs d'agir au-delà de leur niveau de sophistication. Les agresseurs pourront agir plus rapidement et découvrir de nouvelles vulnérabilités du jour zéro, même dans les piles complexes. La manière dont nous priorisons et construisons des mesures préventives doit changer pour faire face à cette menace.

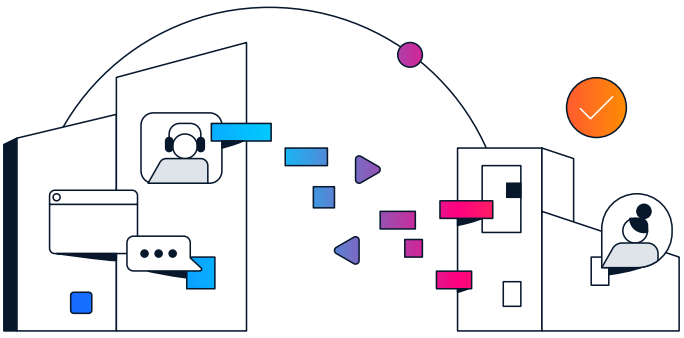
L'adaptation de Cisco pour sécuriser ses produits

Cisco relève le défi de la cyberdéfense de l'ère de l'intelligence artificielle en utilisant ces modèles d'intelligence artificielle avancés pour trouver et corriger les vulnérabilités à une vitesse et à une échelle qui étaient auparavant impossibles, tout en accélérant le développement de produits de sécurité qui peuvent se défendre contre des adversaires utilisant l'intelligence artificielle. Au-delà de la découverte des vulnérabilités et du développement de produits, nous évoluons aussi dans la manière de créer et de valider les logiciels.

Cela comprend la mise à jour de nos modèles de menace pour tenir compte des adversaires assistés par l'intelligence artificielle, l'intégration de scénarios de l'ère de l'intelligence artificielle dans nos exercices d'équipe rouge et, au bout du compte, le dépassement des tactiques, des techniques et des procédures traditionnelles pour soumettre nos produits à des tests de contrainte par rapport aux performances réelles de ces modèles.

Alors que les agents de codage de l'intelligence artificielle font partie intégrante des flux de travail du développement de logiciels, il est essentiel de s'assurer que ces agents produisent un code sécurisé par défaut. Cisco a récemment fait don du projet CodeGuard à la Coalition for Secure AI (CoSAI). Le projet CodeGuard fournit un cadre de sécurité à code source ouvert et indépendant du modèle qui intègre des pratiques de sécurité par défaut directement dans les flux de travail des agents de codage de l'intelligence artificielle.

CodeGuard intègre des compétences et des règles de sécurité qui guident les agents d'intelligence artificielle afin de prévenir les vulnérabilités courantes lors de la génération et de la vérification du code. Cisco recommande aux entreprises d'adopter des cadres comme CodeGuard pour s'assurer que la même accélération de l'intelligence artificielle utilisée pour écrire du code n'introduise pas involontairement des vulnérabilités que des agresseurs utilisant l'intelligence artificielle exploiteront.



Pour renforcer ses capacités défensives, Cisco a lancé Foundry Security Spec, un outil de test à code source ouvert conçu pour aider les entreprises à mettre en place un système agentif d'évaluation de la sécurité. Cette caractéristique éprouvée permet aux équipes de créer un système d'intelligence artificielle qui

répond à leurs besoins uniques en matière d'environnement et de sécurité. En fournissant un cadre commun pour l'évaluation de la sécurité, nous aidons le secteur à créer des systèmes plus fiables et plus sécurisés à grande vitesse.

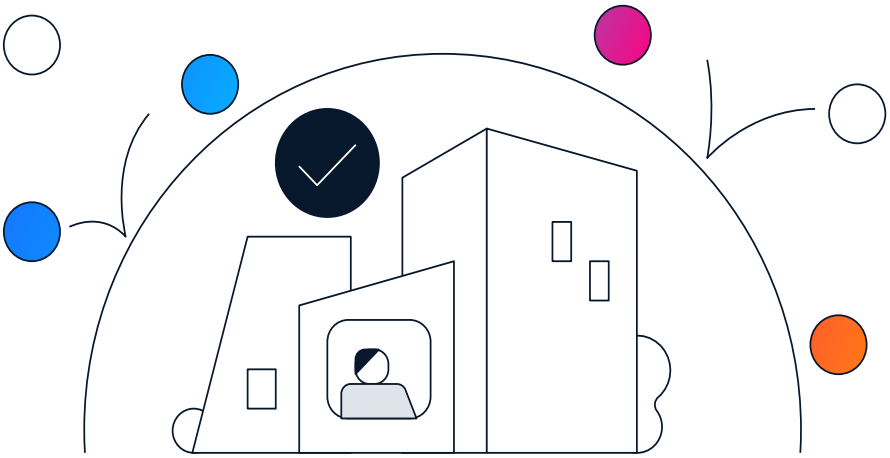
En parallèle, nous mettons ces capacités en œuvre grâce à notre initiative d'infrastructure résiliente, qui met l'accent sur les principes de sécurité par défaut et de sécurité dès la conception, le renforcement proactif de l'infrastructure, l'application rigoureuse de correctifs et la gestion du cycle de vie, ainsi que la suppression systématique des fonctionnalités et des protocoles non sécurisés dans tous les produits Cisco.

Cela comprend le renforcement des configurations par défaut, l'amélioration de la journalisation et de la surveillance pour une télémétrie de sécurité plus riche, et la modernisation de l'authentification des appareils grâce à des protocoles et un chiffrement plus solides, le tout dans le but de réduire la surface d'attaque et d'aider les clients à anticiper les menaces de demain et à y résister. Ces efforts combinés témoignent de l'engagement de Cisco non seulement pour réagir aux menaces émergentes de l'ère de l'intelligence artificielle, mais aussi pour les garder à l'esprit et aider nos clients à bâtir une base numérique plus résiliente.

Notre utilisation précoce des modèles d'intelligence artificielle exploratrice indique que l'ancien modèle d'« une CVE par vulnérabilité » atteint un point de rupture. Étant donné que la découverte automatisée entraîne une augmentation exponentielle des bogues identifiés, le traitement de chaque faille mineure comme un dossier de divulgation individuel engorge l'écosystème de sécurité et retarde activement la mise à jour des logiciels. Notre étoile polaire est de fournir aux clients des renseignements exploitables et non des données volumineuses. En évoluant vers un modèle de divulgation consolidé, selon lequel les vulnérabilités graves sont hiérarchisées et les correctifs mineurs sont intégrés aux cycles de versions standard, nous pouvons accélérer les décisions en matière de correctifs. Cette approche simplifiée refuse aux auteurs de menaces les feuilles de route détaillées dont ils ont besoin pour militariser l'intelligence artificielle contre notre infrastructure.

Pour nous défendre contre les menaces modernes, nous devons privilégier l'action plutôt que l'administration. L'approche traditionnelle qui consiste à attribuer une CVE individuelle à chaque problème mineur crée une « taxe sur les vulnérabilités » qui ralentit les mises à niveau et épuise les équipes de sécurité. Nous croyons que l'avenir de la divulgation doit se concentrer sur les résultats : guider les clients vers des mesures rapides d'atténuation et de mise à niveau. Le secteur a besoin d'un programme de CVE solide qui peut passer à ce nouveau niveau de découverte et de divulgation des failles de sécurité.

Ces efforts témoignent de l'engagement de Cisco à réagir aux menaces émergentes de l'ère de l'intelligence artificielle, à les prévoir et à aider nos clients à établir une base numérique plus résiliente.



Défendre notre entreprise

Nous appliquons aussi ces principes à notre propre environnement d'entreprise. Les recommandations décrites ci-dessous ne sont pas conceptuelles. Elles reflètent la même approche que Cisco adopte à l'interne pour se défendre contre les menaces à l'ère de l'intelligence artificielle. Qu'il s'agisse d'accélérer les cycles de correctifs ou d'éliminer les systèmes en fin de vie, de déployer la recherche de menaces assistée par l'intelligence artificielle et d'appliquer le privilège minimal pour les agents d'intelligence artificielle, nous mettons en œuvre ces conseils dans notre propre infrastructure.

Nos recommandations

Pour répondre efficacement aux capacités d'accélération activées par les modèles d'intelligence artificielle avancés, **les entreprises doivent adopter une approche équilibrée qui renforce les pratiques de sécurité fondamentales tout en modernisant leur architecture préventive**. Si le cadre des menaces évolue rapidement, de nombreuses attaques réussies exploitent encore des faiblesses bien connues. Le renforcement des contrôles principaux demeure l'une des actions les plus percutantes que les responsables de la sécurité puissent prendre.

Les entreprises doivent **accorder la priorité aux mesures fondamentales, comme l'authentification résistante à l'hameçonnage, la vérification renforcée de l'identité, l'accès assorti de privilèges minimaux (y compris les agents d'intelligence artificielle) et les architectures à vérification systématique**. Une gestion cohérente des correctifs, une visibilité complète des ressources et une gestion disciplinée de la configuration sont essentielles afin de réduire les vulnérabilités exploitables. Ces contrôles constituent la référence pour la résilience et sont essentiels pour limiter la portée et la propagation des attaques traditionnelles et de l'ère de l'intelligence artificielle. Dans bien des cas, l'amélioration de l'exécution de ces principes fondamentaux permettra de réduire les risques plus immédiatement que le déploiement seul de nouvelles technologies.

Dans le même temps, les organisations doivent adopter une position agressive afin d'éliminer les risques structuraux. **Les appareils ou les logiciels qui ne peuvent pas faire l'objet d'un correctif, d'une mise à niveau ou d'une prise en charge doivent être systématiquement retirés et remplacés par des plateformes modernes.** Les systèmes modernes intègrent des protections avancées comme des mécanismes de sécurité de la mémoire et des mesures d'atténuation des exploits qui augmentent considérablement la difficulté d'exploiter les vulnérabilités.

Même lorsque des vulnérabilités existent, ces protections ralentissent les agresseurs et réduisent la probabilité d'une exploitation réussie. La création d'environnements flexibles, évolutifs en continu et conçus pour une correction rapide est désormais une exigence critique, particulièrement pour les services Internet, où il y aura très peu de temps entre la divulgation et l'exploitation de masse.

Mais il ne suffit pas de renforcer les bases et de moderniser les infrastructures. La vitesse des attaques fondées sur l'intelligence artificielle réduira l'espace entre la découverte et l'exploitation des vulnérabilités en quelques minutes ou en quelques secondes. Les modèles traditionnels fondés uniquement sur la détection et l'intervention ne sont plus adéquats lorsqu'ils sont utilisés seuls. **Les défenseurs doivent faire évoluer leur modèle opérationnel afin qu'il corresponde à la vitesse, à l'échelle et à l'adaptabilité des menaces de l'ère de l'intelligence artificielle.** Cela comprend un investissement dans la détection de la vitesse de la machine, le triage et le confinement automatisés, ainsi que la surveillance continue de l'identité et de l'activité des données. Le but est de réduire la dépendance à l'égard des interventions manuelles et de permettre de répondre plus rapidement et de manière plus cohérente aux menaces à niveau de confiance élevé.

Cette évolution **nécessite également un passage à la prévention active intégrée.** Plutôt que de se fier exclusivement à la collecte de données télémétriques et à

Équilibrer les contrôles de base et les capacités de prévention adaptatives en temps réel

Renforcer les principes de base

Une authentification multifacteur résistante à l'hameçonnage, la vérification systématique, les privilèges minimaux (y compris les agents d'intelligence artificielle), une gestion disciplinée des correctifs et une visibilité des actifs complète.

Éliminer le risque structurel

Retirez les systèmes en fin de vie. Remplacez-les par des plateformes modernes offrant une sécurité de la mémoire et des mesures d'atténuation des exploits. Conçu pour permettre une évolutivité en continu.

Automatiser à grande vitesse

Investissez dans la détection, le triage et le confinement automatisés. Les modèles de réponse manuelle ne peuvent pas correspondre à la vitesse d'attaque basée sur l'intelligence artificielle.

Intégrer la prévention active

Placez des protections dans la charge de travail, l'appareil et le chemin du trafic : contrôles de temps d'exécution eBPF, application en ligne, boucliers contre les exploits pouvant être mis à jour de manière indépendante.

Utiliser l'intelligence artificielle pour la défense

Utilisez l'intelligence artificielle pour la recherche de menaces, les tests de conformité, les jumeaux numériques et la validation, ce qui réduit les cycles de déploiement de quelques mois à quelques jours.

l'analyse après les événements, les organisations devraient placer des protections directement dans la charge de travail, l'appareil et le chemin du trafic, afin de permettre aux contrôles de sécurité d'agir en temps réel. Par exemple, il peut s'agir de mécanismes d'application en ligne, de protections du temps d'exécution utilisant des technologies comme eBPF pour une visibilité et un contrôle de bas niveau, et de boucliers contre les exploits pouvant être mis à jour indépendamment qui peuvent répondre aux menaces émergentes sans nécessiter de mises à niveau complètes du système. Ces capacités doivent être conçues pour évoluer rapidement, avec la possibilité de mettre à jour les protections indépendamment des grands cycles d'actualisation logiciel ou matériel.

Les entreprises devraient également **exploiter les capacités de l'intelligence artificielle pour leur propre défense**. La recherche constante des menaces internes, aidée par les mêmes modèles efficaces que ceux utilisés par les adversaires, sera une capacité clé pour les défenseurs performants. Les tests de conformité et d'acceptation basés sur l'intelligence artificielle peuvent remplacer la vérification manuelle qui exige beaucoup de travail par des renseignements automatisés à haute vitesse, générant ainsi des scénarios de tests complexes qui couvrent des cas périphériques souvent omis par les testeurs humains. Dans les environnements à enjeux élevés, les jumeaux numériques optimisés par l'intelligence artificielle peuvent simuler les réseaux de production à grande échelle, vérifiant ainsi que les mises à jour respectent des protocoles de sécurité rigoureux et des références en matière de performance, sans compromettre la stabilité des environnements en direct. L'intégration de l'intelligence artificielle dans les phases d'acceptation et de validation réduit considérablement le goulot d'étranglement du déploiement, **ce qui comprime la transition du code terminé au déploiement sur le terrain de plusieurs mois à quelques jours**.

En fin de compte, la réussite dans ce nouvel environnement nécessite un double objectif : exécuter des contrôles de base avec discipline tout en progressant vers des capacités de sécurité adaptatives, en temps réel et intégrées. Les entreprises qui réduisent de manière dynamique les risques existants, modernisent leur infrastructure, adoptent un état d'esprit fondé sur la présomption et adoptent des modèles de prévention active seront les mieux placées pour gérer la vitesse et l'échelle des menaces générées par l'intelligence artificielle.

Conclusion

Le changement arrive. Les défenseurs doivent examiner de manière consciente l'environnement qu'ils défendent aujourd'hui et commencer à façonner cet environnement afin qu'il puisse croître dans un monde antagoniste à l'ère de l'intelligence artificielle. La sagesse d'hier est toujours importante, mais elle doit être combinée à des capacités préventives modernes et de pointe, des réseaux avec une visibilité exceptionnelle et une utilisation appropriée des agents d'intelligence artificielle pour aider les humains à sécuriser les environnements qu'ils défendent.