RAPPORT

La sécurité de l'IA 2026

Observations issues de plus de 625 000 environnements clients à travers le monde

SOPHOS

SOPHOS | X-OPS

Avant-propos de John Peterson

En mai 2026, Sophos a découvert qu'un attaquant menait e qui s'apparentait à une opération de développement logiciel au sein du réseau d'un client. Une douzaine d'agents d'IA, coordonnés par le biais d'un assistant de codage commercial, écrivaient et testaient des malwares contre les protections Endpoint de Sophos, CrowdStrike et Windows Defender, chacun sur sa propre machine virtuelle. L'opération a produit près de 80 modules et plus de 70 techniques d'évasion, chaque résultat ayant été intégré au système de contrôle de version et amélioré lors de l'itération suivante. Le même opérateur a ensuite utilisé ces outils pour déployer un ransomware et voler des données.

Bien que les méthodes elles-mêmes ne soient pas nouvelles, les agents ont réussi à réduire des semaines d'itérations manuelles à seulement quelques jours d'itérations automatisées, et c'est précisément ce changement qui fait l'objet du présent rapport. L'IA modifie la rapidité et la structure des opérations de sécurité plus qu'elle ne modifie les principes fondamentaux des intrusions. Les attaquants continuent d'avoir besoin d'un accès initial, de se déplacer latéralement et d'exfiltrer les données via des canaux observables. Ce qui a changé, c'est la vitesse à laquelle l'attaque se produit.

Dans ce rapport, Sophos démontre cette réalité à l'aide de données concrètes.

Nos analystes MDR (Managed Detection and Response) et IR (Incident Response) investiguent des milliers d'incidents par an. Notre Counter Threat Unit (CTU) surveille les forums clandestins dans plusieurs langues. Nos chercheurs des SophosLabs analysent des échantillons de malwares à grande échelle. Notre équipe Sophos AI développe des techniques innovantes axées sur la défense. Et nous disposons d'une télémétrie relative aux terminaux, aux pare-feux, à la messagerie, au Cloud, au réseau et à l'identité provenant de plus de 625 000 clients.

Cette ampleur, cette portée et cette expertise nous offrent un point de vue que peu d'acteurs possèdent sur l'évolution rapide de l'intersection entre IA et cybersécurité. Nous entendons utiliser cette perspective pour aider le secteur à garder une longueur d'avance et garantir la protection de nos clients.

J. P. Peterson, CTO, Sophos

Résumé et principales conclusions

Ce rapport s'appuie sur les dossiers traités par Sophos MDR et Sophos IR, les analyses de SophosLabs, les renseignements de Sophos CTU, ainsi que sur les observations relatives aux postes et aux réseaux d'une base de clients de plus de 625 000 organisations à travers le monde. Une tendance se dégage de tout cela : L'IA permet aux attaquants et aux équipes de défense de réaliser des opérations courantes beaucoup plus rapidement. Même si les attaques totalement inédites restent limitées, sans pour autant être inexistantes, elles pourraient augmenter avec la montée en puissance des capacités.

2026 marque un point d'inflexion. Les modèles frontière et open-weight sont désormais suffisamment performants pour permettre une délégation effective de tâches d'ingénierie, tandis que la baisse des coûts d'inférence modifie l'économie du secteur pour les deux camps. La question pratique n'est plus de savoir si l'IA sera utilisée, mais d'identifier où les systèmes frontière, coûteux, se justifient, où des modèles open-weight, plus économiques, suffisent, et comment organiser les flux de travail sans accroître l'exposition au risque.

Les constats suivants constituent les piliers du rapport :

1. **L'IA accélère la rapidité des attaques, mais n'introduit pas de nouveaux types d'attaques.** L'acteur malveillant que Sophos suit sous l'identifiant STAC6994 a utilisé des agents d'IA pour itérer des techniques de contournement des solutions EDR à un rythme qui aurait demandé des semaines à un développeur humain. Les analystes de Sophos CTU ont ensuite confirmé que l'opérateur derrière STAC6994 avait mené des opérations de déploiement de ransomware et de détournement de données.
2. **La couche d'identifiants et d'identités associée aux services d'IA d'entreprise est désormais une cible privilégiée pour les attaques de collecte de données.** L'incident Salesloft/Drift a démontré comment des jetons OAuth de chatbot peuvent se traduire directement par une compromission de l'environnement Salesforce. Par ailleurs, les analystes de Sophos CTU ont observé des cybercriminels déclarer sur des forums clandestins que des prompts IA et des données générées peuvent être capturés comme dommages collatéraux lors d'attaques.
3. **L'économie souterraine absorbe l'IA comme infrastructure.** Sophos a constaté que certains cybercriminels vendaient des clés API obtenues par courtage pour Claude, ChatGPT et Grok. Sur les forums, de nouveaux canaux se sont formés autour de l'ingénierie de prompts IA, des techniques de jailbreak et des processus de développement de malware. Certains profils recrutent des ingénieurs en prompt IA. D'autres restent ouvertement sceptiques quant au fait que l'IA changera leurs opérations. La dynamique d'adoption ressemble moins à une transformation soudaine qu'à l'intégration progressive d'un nouvel outil dans des workflows criminels existants.
4. **Les attaques de la chaîne d'approvisionnement ciblent désormais spécifiquement l'infrastructure de développement de l'IA.** L'attaque visant l'extension Nx Console de VS Code a abouti à la distribution d'un malware conçu pour dérober des

identifiants. Les analystes SophosLabs ont examiné de fausses pages de téléchargement de Claude diffusant un nouveau RAT, ainsi qu'une campagne ClickFix qui détournait des URL légitimes de chats partagés de ChatGPT afin de déployer l'infostealer MacSync.

5. **L'ingénierie sociale assistée par l'IA est passée du stade expérimental au stade opérationnel.** Sophos CTU a documenté des annonces publiées sur des forums clandestins proposant des voice bots, des profils générés par IA destinés aux opérations d'escroqueries sentimentales, ainsi que des services de création d'images de profil de synthèse. La contribution de l'IA réside dans sa scalabilité et la qualité linguistique, appliquées aux mêmes techniques de tromperie qui fonctionnaient déjà avant l'essor de l'IA générative.

6. **Les failles sont exploitées de plus en plus rapidement avant même qu'elles ne soient corrigées.** La Binding Operational Directive (directive opérationnelle contraignante) 26-04 de la CISA (Cybersecurity and Infrastructure Security Agency) publiée le 10 juin 2026 a réduit le délai obligatoire pour l'application des correctifs à trois jours pour les failles présentant les plus grands risques, citant explicitement l'IA comme un facteur réduisant la fenêtre entre divulgation et exploitation. Le premier test réel de la directive est arrivé presque immédiatement : CVE-2026-10520 (Ivanti Sentry) a été exploité dans les 24 heures suivant la publication de la preuve de concept et ajouté au catalogue KEV le 12 juin. CVE-2026-42208 (LiteLLM) a été exploité dans les 36 heures, les attaquants ciblant des bases de données contenant des clés de fournisseurs LLM en amont.

7. **Prouver l'utilisation de l'IA dans une attaque reste opérationnellement difficile.** Le cas de STAC6994 était atypique, dans la mesure où le framework de développement se trouvait encore sur un appareil accessible par Sophos pour examen, offrant une visibilité rare sur les outils utilisés par l'opérateur. Dans un grand nombre d'incidents, l'usage de l'IA demeure indétectable au niveau de la télémétrie. La prévalence réelle des attaques assistées par IA est probablement plus élevée que ce que tout éditeur peut prouver directement.

8. **Les modèles de raisonnement IA accélèrent les investigations.** Les modèles de raisonnement IA réduisent des heures d'investigation réalisées par des analystes à seulement quelques minutes pour des schémas d'attaque connus. Ils ne remplacent pas les systèmes de détection comportementale capables de repérer une menace isolée au milieu de milliards d'événements. L'extrême déséquilibre des classes, la dépendance à des bases de référence adaptées à chaque environnement et la vitesse d'évolution des techniques adverses créent des contraintes statistiques et opérationnelles qui nécessitent une ingénierie spécialisée.

Partie 1 :
L'IA dans la chaîne d'attaque

Taxonomie Sophos X-Ops des menaces liées à l'IA

Sophos X-Ops distingue deux grandes catégories de menaces liées à l'IA : l'usage malveillant de l'IA (les menaces contre lesquelles nous protégeons les utilisateurs) et le ciblage malveillant de l'IA (l'IA que nous protégeons). La première catégorie couvre les artefacts générés par l'IA, les capacités opérationnelles augmentées par l'IA et les intrusions orchestrées par l'IA. La seconde couvre les compromissions initiées par des agents, l'usurpation d'identité de logiciels d'IA, les LLM empoisonnés et les attaques contre les modèles eux-mêmes. Les deux catégories exigent des approches distinctes : l'usage malveillant de l'IA pose un défi de productivité et de scalabilité, tandis que le ciblage malveillant de l'IA soulève des questions de confiance, de chaîne d'approvisionnement et de gouvernance. Cette taxonomie complète des cadres existants tels que MITRE ATLAS, la taxonomie des modes de défaillance des agents de Microsoft, l'AI Risk Repository du MIT et le NIST AI 100-2.

Figure 1 : Aperçu de la taxonomie des menaces liées à l'IA

Une échelle d'autonomie

Au niveau le plus minimal, les attaques générées par l'IA impliquent un humain utilisant une IA générative pour produire un artefact, puis le déployant manuellement. Un acteur malveillant ciblant des organisations gouvernementales mexicaines a utilisé Claude Code et GPT pour générer des scripts. Des discussions interceptées entre le groupe de ransomware The Gentlemen ont révélé des applications similaires. Et, dans un cas récent de ransomware DragonForce investigué par l'équipe Sophos IR, son auteur a produit ce que nous estimons très probablement être un rapport généré par l'IA au cours des négociations avec l'organisation ciblée, incluant une analyse approfondie des données exfiltrées, notamment des données personnelles d'identification (PII) et des risques juridiques associés. L'analyse générée par IA a été utilisée pour tenter d'accroître la pression sur l'organisation et la convaincre de payer la rançon.

À un niveau d'autonomie supérieur, les attaques augmentées par IA reposent sur un modèle qui améliore une capacité au runtime. LameHug, un malware basé sur Python que le CERT-UA attribue avec une confiance modérée à IRON TWILIGHT (APT28), a interrogé un modèle open-weight hébergé afin de générer dynamiquement, selon l'environnement, des commandes de reconnaissance Windows. Cette approche réduit fortement l'utilité de l'analyse statique et fait du trafic sortant vers des API d'IA/ML l'une des rares surfaces de détection fiables. L'usurpation d'identité assistée par IA représente un autre vecteur très répandu : le clonage de voix et les outils de face-swap sont aujourd'hui employés dans le cadre de fraudes en temps réel, d'usurpations d'identité de cadres supérieurs et de tentatives de contournement du KYC (Know Your Customer).

À l'extrémité la plus éloignée, la divulgation d'Anthropic en novembre 2025 a documenté une campagne soutenue par la Chine utilisant Claude Code pour tenter de compromettre environ trente cibles. Ces attaques restent rares, mais elles réduisent considérablement le temps entre la reconnaissance et l'action, remettant en cause les postulats traditionnels de la défense.

Étude de cas : *STAC6994*

Les analystes de Sophos ont observé un acteur malveillant utilisant l'IA pour tester des tactiques de contournement de la technologie EDR dans un framework de post-exploitation de type « Red Team », découvrant un appareil non autorisé générant en continu des payloads dans l'environnement d'un client. L'attaquant avait provisionné des machines virtuelles depuis Ludus et utilisait l'IDE Cursor avec des agents d'IA pour développer et tester des outils de post-exploitation.

Le framework a exécuté des tests parallèles sur trois environnements distincts : une VM avec Sophos Endpoint Protection, une VM avec CrowdStrike, et une VM dépourvue de technologie EDR utilisée comme environnement témoin. Une quatrième VM fonctionnait comme serveur Sliver C2. Environ 12 agents d'IA fonctionnaient avec des rôles définis.

Un agent, exécutant Claude Opus 4.5, gérait les opérations principales et la définition des règles. D'autres géraient le renforcement de la sécurité opérationnelle (OPSEC), la documentation, les tests de résistance des proxys, le déploiement de VM et les tests d'outils contre les agents EDR. Les commits de code étaient transmis à Git via le Model Context Protocol (MCP).

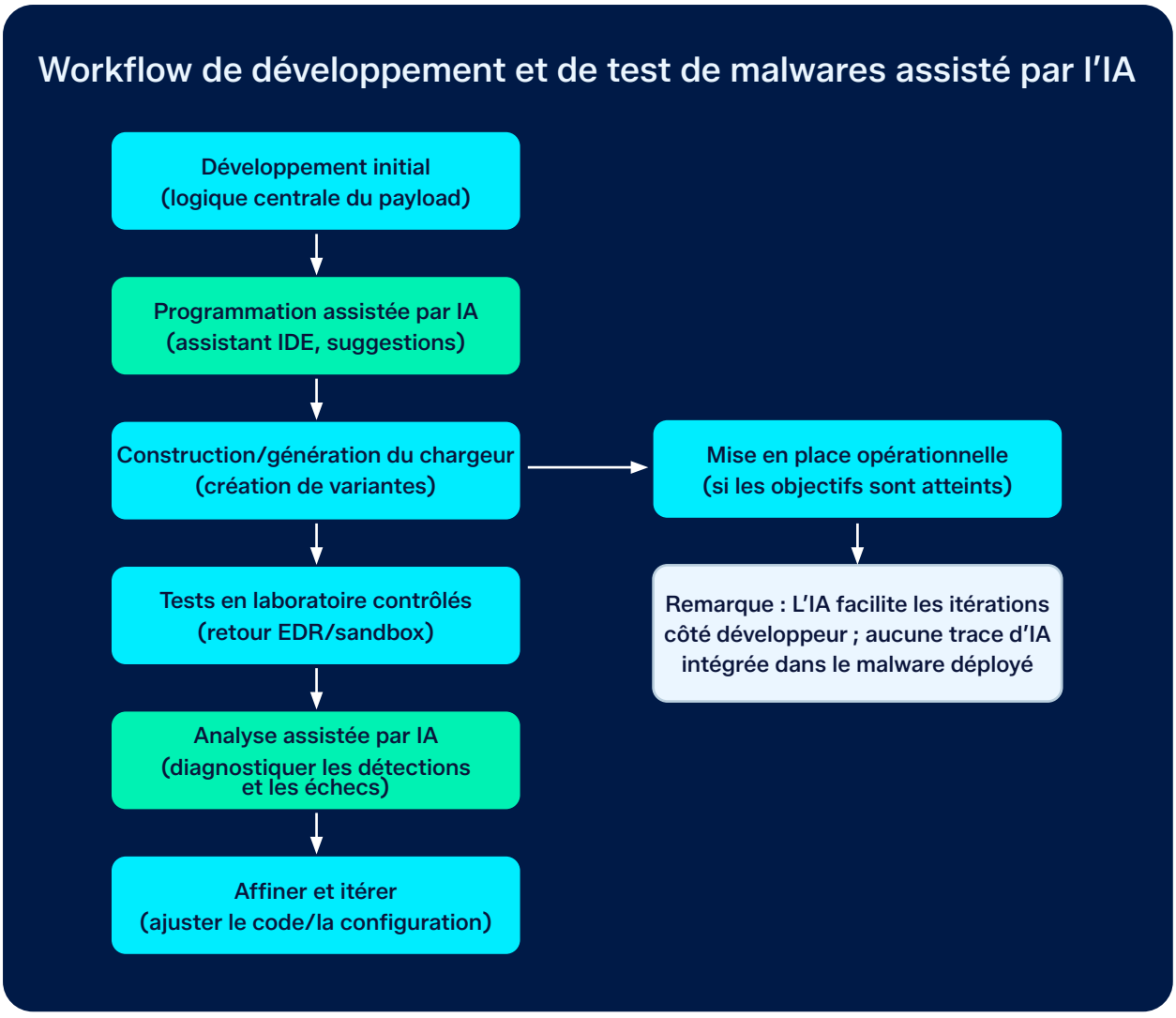


Figure 2 : Schéma montrant le rôle de l'IA dans le flux de travail de développement de malwares

Le playbook suivait une approche systématique. Selon les artefacts récupérés, les agents ont lu des articles de recherche récupérés sur le blog de SpecterOps, en ont extrait des techniques offensives, les ont mappées au frameworkMITRE ATT&CK, ont identifié les étapes et les outils nécessaires pour reproduire chaque technique, ont préparé l'environnement de laboratoire, l'ont exécuté et ont rapporté les conclusions.

Au cœur du dispositif se trouvait un générateur de payloads modulaire en Python, qui enveloppait les payloads bruts dans des couches de chiffrement et de techniques de contournement, produisant des exécutables sur mesure en Rust et en Go. Près de 80 modules testant plus de 70 techniques différentes ont été développés. Le chemin de contournement réel de la technologie EDR était un cycle de test d'ingénierie structuré avec examen humain et itération. L'IA coordonnait le workflow et facilitait l'expérimentation. Les éléments recueillis n'indiquent pas la présence d'IA intégrée dans le malware déployé.

Après diverses itérations, la documentation produite par les agents affirmait une efficacité presque totale face aux agents EDR, mais les analystes de Sophos ont relevé que les éléments disponibles ne confirmaient pas entièrement cette conclusion. Le signal opérationnel clé résidait dans le tempo : l'ensemble du cycle réduisait des semaines de travail à seulement quelques jours.

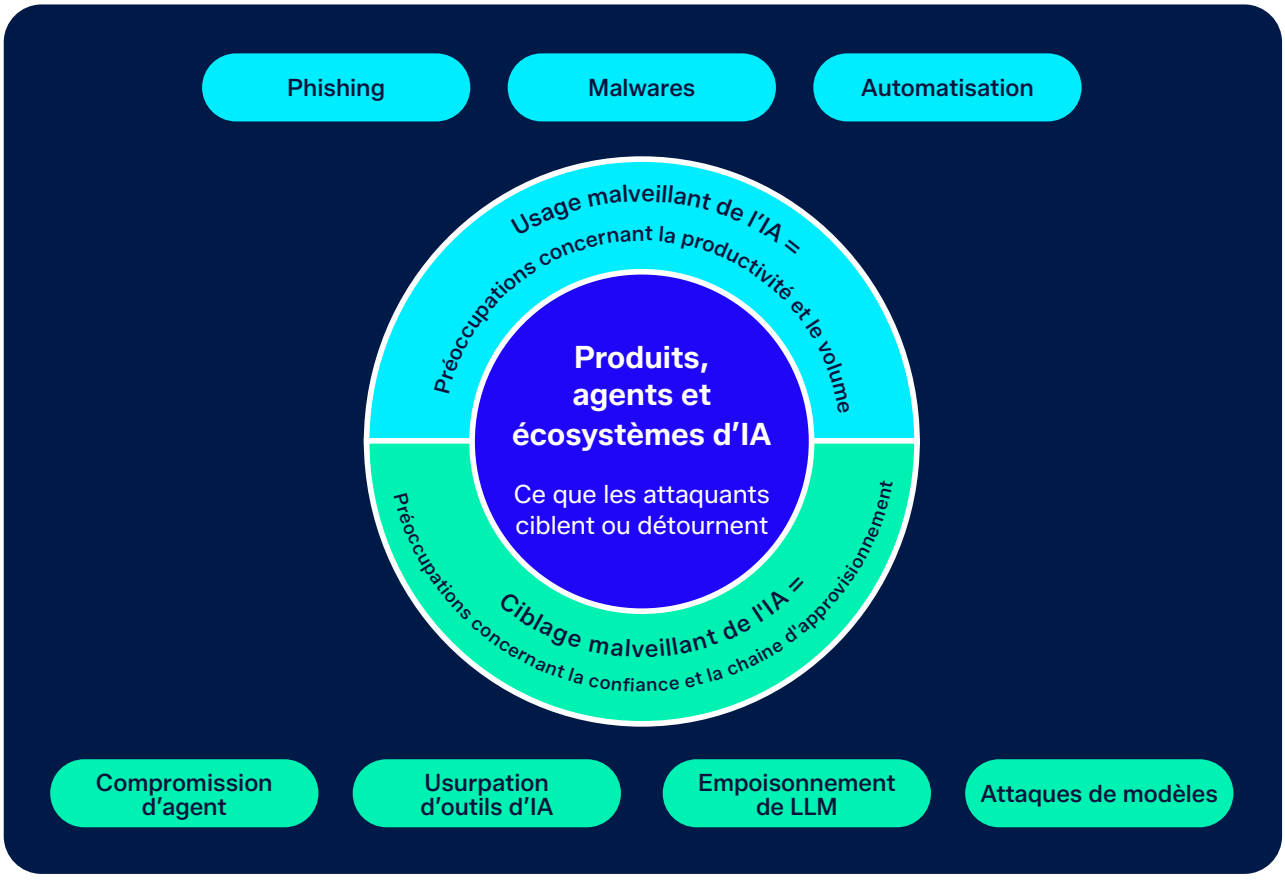
Sophos CTU a confirmé que l'opérateur STAC6994 avait par la suite mené des opérations de déploiement de ransomware et de vol de données, indiquant que le framework était un outil opérationnel utilisé dans des intrusions réelles.

Ciblage malveillant de l'IA

La seconde partie de notre taxonomie couvre les attaques dans lesquelles les produits, agents et écosystèmes d'IA constituent la cible ou la surface d'attaque, plutôt que l'outil. Nous examinerons cela plus en détail dans la Partie 2. La compromission initiée par l'agent est l'un des sous-types les plus préoccupants : un agent de codage qui télécharge un package NPM (Node Package Manager) compromis ou interagit avec un serveur MCP (Model Context Protocol) empoisonné dans le cadre de son travail. Le risque réside dans la réduction du délai entre la publication et l'exécution, en l'absence totale d'un humain dans la boucle pour contrôler le changelog.

Notre taxonomie couvre aussi l'usurpation de logiciels d'IA (des acteurs malveillants profitant d'utilisateurs en quête d'outils d'IA légitimes), l'empoisonnement de LLM (qu'il s'agisse de configurer un modèle sur mesure ou d'injecter délibérément du contenu malveillant ou biaisé pour infléchir son comportement), ainsi qu'un ensemble d'attaques ciblant directement les modèles : extraction de modèle, inversion des données d'entraînement, exemples adversariaux et l'inférence d'appartenance.

Traiter séparément ces deux catégories de haut niveau offre une vision plus claire de la menace. L'usage malveillant de l'IA est fondamentalement un problème de productivité et de volume : les piles de détection existantes en interceptent une partie, mais la difficulté réside dans l'échelle et la vitesse d'itération. Le ciblage malveillant de l'IA relève, lui, d'un enjeu de confiance et de chaîne d'approvisionnement, plus efficacement traité par des politiques de sécurité, des cadres de conception sécurisée (Secure-by-Design) et des initiatives similaires.



Adoption clandestine et scepticisme

Sophos suit l'évolution des attitudes à l'égard de l'IA générative sur les forums criminels depuis 2023, avec une étude de suivi réalisée en 2025, et une veille continue depuis. La situation a considérablement évolué. Les acteurs malveillants achètent des clés API courtées pour ChatGPT, Claude et Grok, et les forums ont vu apparaître des canaux dédiés à l'IA où des profils partagent des modèles de prompts et des techniques de jailbreak. Les chercheurs de Sophos CTU ont relevé qu'un recruteur clandestin connu, auparavant observé en train d'embaucher des développeurs blockchain et du personnel de centres d'appels pour des opérations de phishing, cherchait désormais à recruter spécifiquement un ingénieur en prompts IA, externalisant l'expertise IA comme les acteurs malveillants externalisent le développement de malware ou l'accès initial.

Sophos CTU a également observé des annonces pour des voice bots IA dédiés au vishing et à la fraude par appel, avec des retours d'utilisateurs indiquant que ces bots savent reproduire de façon convaincante le ton, la cadence et les dynamiques conversationnelles. Le profil « HackingRealm » a promu un service "AI OnlyFans Models" destiné à l'escroquerie sentimentale et à l'ingénierie sociale, générant des images de profil de synthèse et du contenu conversationnel pour créer des profils à grande échelle sur les plateformes de rencontre et les réseaux sociaux, assorti d'un site web dédié et d'une page de commentaires affichant des avis positifs.

Des outils commercialisés comme « pilotés par l'IA » commencent à apparaître : Leak Bazaar (triage de données volées optimisé par ML) et une version mise à jour de Cobalt Strike intégrant une REST API et un serveur MCP. Les adversaires ajoutent des intégrations de LLM à des workflows familiers. La référence à Claude dans l'annonce de Cobalt Strike sert en partie de signal de modernité, mais reflète aussi la manière dont l'intégration de LLM devient un facteur de différenciation sur les marchés criminels, même lorsqu'elle apporte surtout de la commodité plutôt qu'un tradecraft perfectionné.

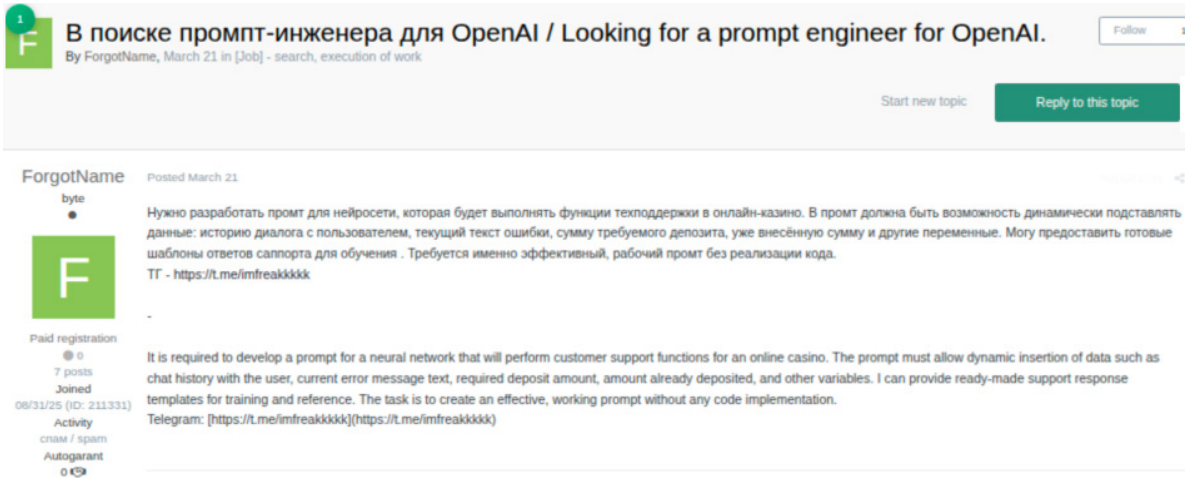
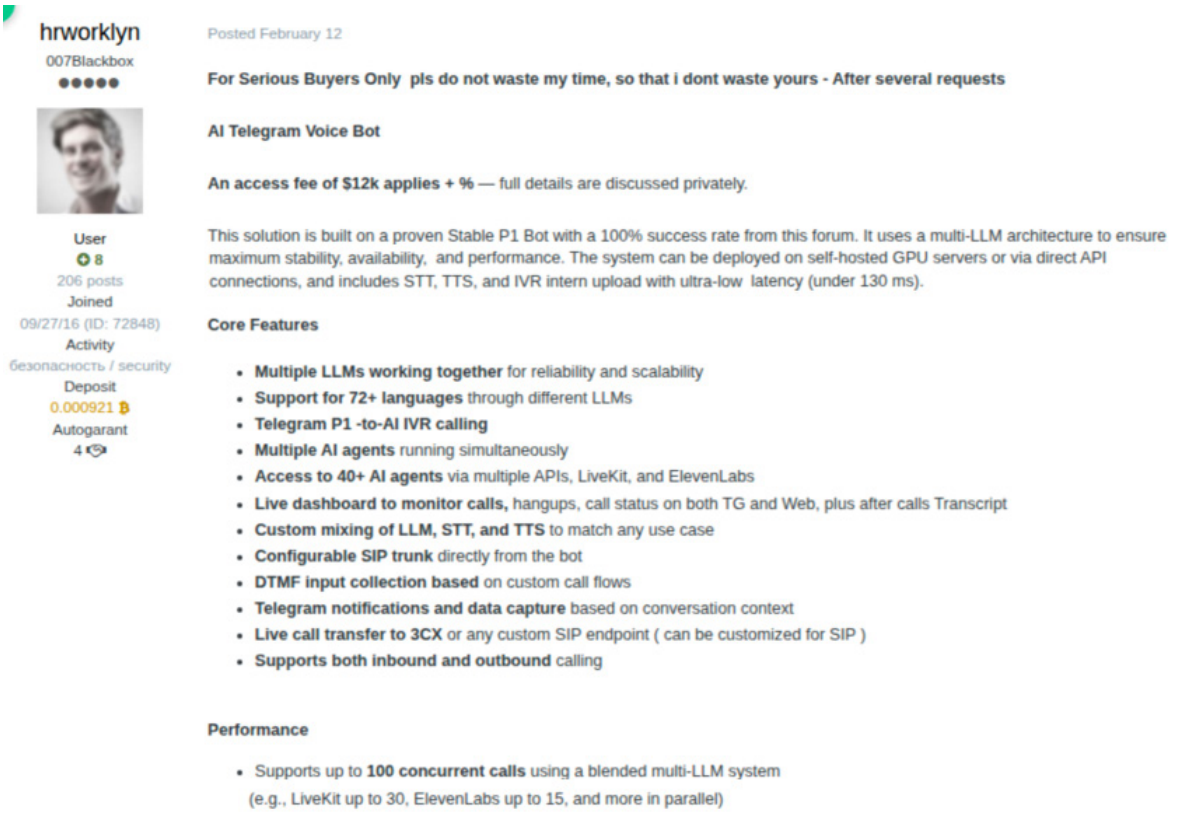


Figure 3 : Annonce de recrutement pour un ingénieur de prompt OpenAI



hrworklyn
007Blackbox
●●●●●

User
8
206 posts
Joined
09/27/16 (ID: 72848)
Activity
безопасность / security
Deposit
0.000921 ₿
Autogarant
4

Posted February 12

For Serious Buyers Only pls do not waste my time, so that i dont waste yours - After several requests

AI Telegram Voice Bot

An access fee of $12k applies + % — full details are discussed privately.

This solution is built on a proven Stable P1 Bot with a 100% success rate from this forum. It uses a multi-LLM architecture to ensure maximum stability, availability, and performance. The system can be deployed on self-hosted GPU servers or via direct API connections, and includes STT, TTS, and IVR intern upload with ultra-low latency (under 130 ms).

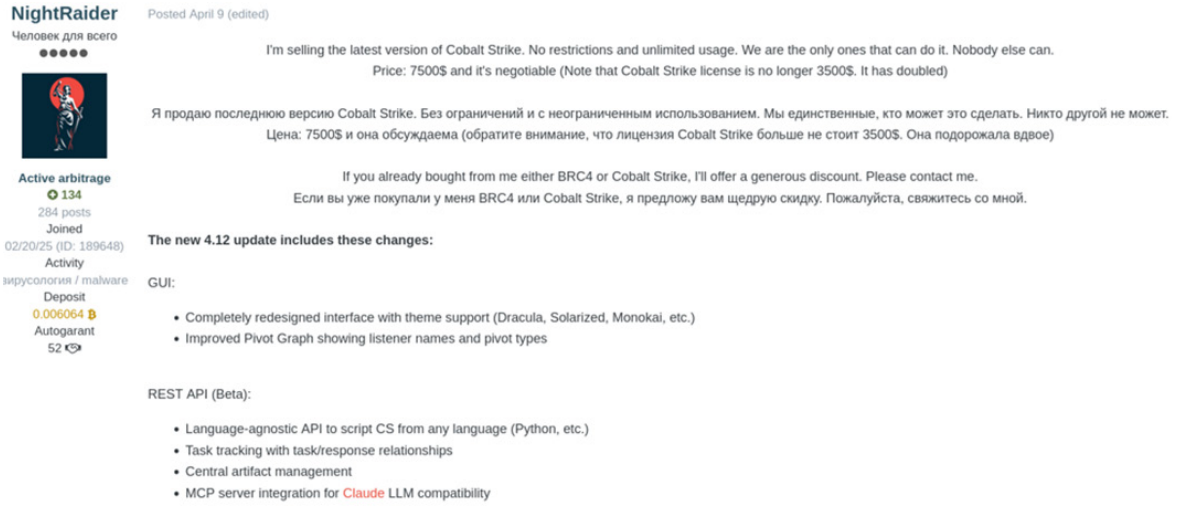
Core Features

- **Multiple LLMs working together** for reliability and scalability
- **Support for 72+ languages** through different LLMs
- **Telegram P1 -to-AI IVR calling**
- **Multiple AI agents** running simultaneously
- **Access to 40+ AI agents** via multiple APIs, LiveKit, and ElevenLabs
- **Live dashboard to monitor calls,** hangups, call status on both TG and Web, plus after calls Transcript
- **Custom mixing of LLM, STT, and TTS** to match any use case
- **Configurable SIP trunk** directly from the bot
- **DTMF input collection based** on custom call flows
- **Telegram notifications and data capture** based on conversation context
- **Live call transfer to 3CX** or any custom SIP endpoint (can be customized for SIP)
- **Supports both inbound and outbound** calling

Performance

- Supports up to **100 concurrent calls** using a blended multi-LLM system (e.g., LiveKit up to 30, ElevenLabs up to 15, and more in parallel)

Figure 4 : Annonce pour un bot vocal IA sur Telegram



NightRaider
Человек для всего
●●●●●

Active arbitrage
134
284 posts
Joined
02/20/25 (ID: 189648)
Activity
вирусология / malware
Deposit
0.006064 ₿
Autogarant
52

Posted April 9 (edited)

I'm selling the latest version of Cobalt Strike. No restrictions and unlimited usage. We are the only ones that can do it. Nobody else can.
Price: 7500$ and it's negotiable (Note that Cobalt Strike license is no longer 3500$. It has doubled)

Я продаю последнюю версию Cobalt Strike. Без ограничений и с неограниченным использованием. Мы единственные, кто может это сделать. Никто другой не может.
Цена: 7500$ и она обсуждаема (обратите внимание, что лицензия Cobalt Strike больше не стоит 3500$. Она подорожала вдвое)

If you already bought from me either BRC4 or Cobalt Strike, I'll offer a generous discount. Please contact me.
Если вы уже покупали у меня BRC4 или Cobalt Strike, я предложу вам щедрую скидку. Пожалуйста, свяжитесь со мной.

The new 4.12 update includes these changes:

GUI:

- Completely redesigned interface with theme support (Dracula, Solarized, Monokai, etc.)
- Improved Pivot Graph showing listener names and pivot types

REST API (Beta):

- Language-agnostic API to script CS from any language (Python, etc.)
- Task tracking with task/response relationships
- Central artifact management
- MCP server integration for Claude LLM compatibility

Figure 5 : Membre d'un forum cybercriminel annonçant une mise à jour de Cobalt Strike.

Mais si l'on voit clairement des signes d'expérimentation active et d'adoption, la communauté, du moins sur les forums surveillés par Sophos CTU, n'est pas uniformément enthousiaste. Reprenant les constats précédents de Sophos sur les attitudes envers l'IA générative en 2023 et 2025, Sophos CTU a observé des profils exprimant la crainte que l'IA réduise les opportunités de travail, en particulier pour le développement de malware et la rédaction de scripts. D'autres rejettent l'IA d'emblée et encouragent la dépendance aux compétences humaines.

Ingénierie sociale améliorée par l'IA et deepfakes

L'IA modifie la rapidité avec laquelle les campagnes d'ingénierie sociale se développent, leur capacité à paraître convaincantes dans différentes langues et leur coût de production. Le rapport M-Trends 2026 de Mandiant, par exemple, a décrit une évolution du phishing générique vers une ingénierie sociale personnalisée et fondée sur les LLM, tandis que le phishing vocal (vishing) est devenu le deuxième vecteur initial d'infection le plus courant (11 %), tandis que le phishing par email est tombé à 6 %. Le rapport sur les menaces MSP 2026 de ConnectWise a confirmé l'utilisation active de la fraude par deepfake et des campagnes de phishing générées par LLM, notant que l'IA rendait les tactiques établies plus rapides, plus évolutives et plus convaincantes plutôt que de créer de nouvelles catégories d'attaques.

De nombreuses catégories d'acteurs malveillants utilisent les deepfakes comme un outil opérationnel, notamment en Corée du nord, où des agents utilisent des identités générées par l'IA et des deepfakes pour obtenir des emplois à distance dans des entreprises occidentales, leur permettant d'établir un accès interne persistant. Sophos a publié un guide CISO sur les tactiques des travailleurs informatiques nord-coréens dédié qui couvre les indicateurs de détection et les contre-mesures organisationnelles.

Sophos CTU a enquêté sur une escroquerie de type « sha zhu pan » reposant sur une ingénierie sociale exploitant des thèmes liés à l'IA. Une victime basée au Royaume-Uni a été attirée vers une fausse plateforme d'investissement optimisée par l'IA appelée DAIQ Wealth grâce à des mois de « leçons » sur le thème de l'IA sur l'appli de messagerie LINE. Tout un réseau de faux profils a établi une relation avec la victime par le biais de discussions privées, chacune étant gérée par une personne différente, ce qui suggère une main-d'œuvre organisée suivant des scripts prédéfinis. La victime a perdu des centaines de milliers de livres sterling. Lorsqu'elle a exprimé son inquiétude à l'idée de conserver près de 20 000 livres en espèces à son domicile, les escrocs ont organisé l'envoi d'un coursier à son adresse au Royaume-Uni en moins de 24 heures, avec un reçu portant la marque (contrefaite) d'une société financière légitime.

En résumé

Comme nous l'avons vu dans notre étude de cas STAC6994 (le recrutement clandestin d'ingénieurs en prompts), l'IA présentée comme un argument commercial dans la vente de services cybercriminels et les campagnes d'escroqueries axées sur l'IA indiquent que les acteurs malveillants adoptent l'IA pour renforcer et faciliter leurs attaques, tout en continuant d'expérimenter activement son potentiel.

Un tableau similaire se dessine dans l'ensemble du secteur. Claude a aidé à mener des attaques contre les réseaux du gouvernement mexicain. Le rapport sur les risques et la résilience de l'IA de Mandiant décrit deux familles de malwares documentées qui interrogent des LLM pendant leur exécution : PROMPTFLUX, un dropper VBScript expérimental utilisant l'API Gemini pour une auto-modification juste-à-temps, et PROMPTSTEAL, attribué à IRON TWILIGHT (APT28), marquant la première observation confirmée d'un malware soutenu par un État interrogeant un LLM lors d'opérations réelles contre l'Ukraine. Le Verizon DBIR a documenté VoidLink (un framework de malwares assemblé par un agent d'IA en six jours) et PromptLock (un ransomware optimisé par l'IA découvert pour la première fois par ESET en août 2025, bien qu'il ait été révélé plus tard qu'il s'agissait très probablement d'une preuve de concept universitaire, et non d'un échantillon malveillant actif détecté dans le monde réel).

Cependant, des campagnes d'intrusion véritablement autonomes et entièrement pilotées par l'IA n'ont, à ce stade, jamais été confirmées à grande échelle. Le Rapport 2026 sur les ransomwares de GuidePoint GRIT a explicitement déclaré que les inquiétudes concernant les « super-affiliés déployant des robots de rançon entièrement autonomes » ont été exagérées, et que l'utilisation de l'IA/LLM reste rudimentaire parmi les acteurs malveillants moins matures. L'évaluation AIM3 de Recorded Future a placé la majorité des malwares IA observés aux niveaux de maturité 1 à 3 de son AI Malware Maturity Model (Experimenting, Adopting, Optimizing), notant l'absence d'exemples confirmés de malwares véritablement « Bring-Your-Own-AI » embarqués, exécutant des modèles locaux sur les hôtes victimes. L'analyse concluait que « les équipes de défense devraient privilégier la surveillance de l'abus de services d'IA légitimes, le renforcement des contrôles existants et le mappage des menaces aux niveaux AIM3, réagir de manière excessive à des scénarios dignes de la science-fiction ».

Le cas GTG-1002 documenté par Anthropic (où Claude Code avait automatisé 80 à 90 % d'une campagne de vol de données multi-cibles visant environ 30 entités) est ce qui s'approche le plus d'une opération autonome dans le domaine public, mais il demeure une anomalie plutôt qu'un indicateur de tendance.

Partie 2 :
Le risque lié à l'IA en entreprise

L'IA est déjà intégrée dans l'entreprise : des agents de codage produisent et déploient du code en utilisant des identifiants de développeurs, des assistants agentiques lisent des emails et appellent des API, des systèmes auparavant isolés sont désormais reliés entre eux, et des équipes internes expérimentent des modèles open-weight sur des infrastructures Cloud managées. Les modèles open-source/open-weight sont de plus en plus efficaces et coûtent une fraction du prix des modèles closed-source/closed-weight ; de plus ils permettent aux entreprises de rester propriétaires de leurs inférences et de garder le contrôle de leurs données, même si celles-ci proviennent pour la plupart de la République populaire de Chine. La question de la gouvernance ne consiste plus à savoir s'il faut autoriser ou non son adoption, mais comment la sécuriser avant que la surface d'attaque ne se consolide autour de paramètres par défaut vulnérables.

Même les organisations les plus réticentes à prendre des risques, qui n'utilisent pas massivement l'IA dans des scénarios d'utilisation autorisés, sont probablement confrontées au problème de Shadow AI et, comme nous l'aborderons bientôt, les craintes concernant le Shadow AI et la fuite de données sont fondées. Nous aborderons plus en détail la tolérance au risque, les contrôles et la visibilité dans la partie 4.

Une tendance distincte en 2026 est le ciblage des infrastructures de développement propres à l'IA, et plus particulièrement des relations de confiance que les développeurs établissent autour des outils d'IA.

L'infrastructure de développement IA comme cible d'attaque

Une tendance distincte en 2026 est le ciblage des infrastructures de développement propres à l'IA, et plus particulièrement des relations de confiance que les développeurs établissent autour des outils d'IA. Contrairement à la plupart des menaces liées à l'IA, encore émergentes ou largement confinées à des preuves de concept théoriques, c'est dans ce domaine que les attaques se produisent en ce moment même.

La campagne du ver npm **SANDWORM_MODE** reposait sur au moins 19 packages typosquattés imitant des utilitaires de développement légitimes et des outils de programmation IA. Une fois installés, ces packages déployaient un serveur MCP malveillant et, grâce à une injection de prompt embarquée, détournaient des assistants IA légitimes pour qu'ils récupèrent discrètement des clés SSH et des identifiants Cloud, avant de les exfiltrer sans que l'utilisateur ne s'en aperçoive.

La **compromission de l'extension Nx Console VS Code** a permis de récolter des identifiants provenant de HashiCorp Vault, npm, AWS, GitHub, 1Password et des clés API Anthropic, et semblait s'inscrire dans la campagne plus large TeamPCP/mini-Shai-Hulud. Une extension compromise au sein de l'IDE d'un développeur dispose d'un accès direct aux identifiants et aux répertoires les plus sensibles.

La **fuite du code source de Claude Code d'Anthropic** en mars 2026, où le code a été publié par inadvertance dans un package npm, illustre une autre dimension. Le code aurait contenu environ 1 900 fichiers et 500 000 lignes, incluant des détails sur des fonctionnalités non publiées. Anthropic a déclaré qu'aucune donnée client n'avait été exposée, mais l'analyse du code pourrait faciliter l'identification future de bugs et de vulnérabilités. Les outils d'IA eux-mêmes sont devenus des cibles de renseignement.

La couche de gestion des identités qui relie les services d'IA aux systèmes de l'entreprise crée un risque pour lequel les mécanismes de gouvernance en place ne sont pas conçus.

Notre recommandation : toute organisation disposant d'une équipe de développement devrait considérer ce risque comme prioritaire et immédiat. Les contrôles mis en place reposent en grande partie sur des processus (gestion rigoureuse des dépendances, verrouillage de version et revue systématique des nouvelles bibliothèques open-source avant leur intégration) associés à une surveillance étroite des terminaux des développeurs. (Dans la Partie 4, nous proposons une liste de sept mesures que les équipes de défense mettre en œuvre dès maintenant pour réduire l'ampleur de l'impact du déploiement de l'IA agentique.

Où se concentre l'exposition

La couche de gestion des identités reliant les services d'IA aux systèmes de l'entreprise crée une vulnérabilité que les mécanismes de gouvernance actuels n'ont pas été conçus pour gérer. IBM X-Force a déclaré que plus de 300 000 identifiants ChatGPT étaient apparus à la vente sur le Dark Web en 2025, et qu'un message publié sur un forum en février 2025 évoquait le vol de plus de 20 millions de comptes ChatGPT ; un chiffre toutefois non confirmé. L'incident Salesloft/Drift a démontré un mécanisme concret : des jetons OAuth Drift ont été utilisés pour compromettre plusieurs environnements Salesforce, démontrant ainsi comment l'accès par jeton d'un chatbot peu entraîner directement une compromission du système.

BeyondTrust a signalé une augmentation de 466,7 % des agents d'IA actifs dans les environnements d'entreprise au cours de l'année écoulée. Ces agents introduisent des surfaces d'attaque spécifiques : injection de prompt, invocation malveillante d'outils et saisies de données empoisonnées. La vulnérabilité CVE-2025-32711 de Copilot de Microsoft (EchoLeak), a démontré comment un faible assainissement des données peut permettre une fuite de données zero-click. Lorsque les identités des machines ne disposent pas d'équivalents de l'authentification multifacteur (MFA), qu'elles transportent des secrets de longue durée et qu'elles opèrent avec des privilèges excessifs, elles deviennent une cible viable et attractive.

Les attaquants sont pleinement conscients de ces opportunités et procèdent activement à l'énumération de la surface d'attaque. En janvier 2026, GreyNoise a décrit des campagnes visant à mapper les déploiements de LLM. Des chercheurs ont documenté l'opération Bizarre Bazaar, une campagne où les attaquants ont détourné des terminaux LLM et MCP exposés, les ont validé, puis ont revendu leurs accès via une infrastructure liée à silver.inc.

Le déficit de gouvernance

Dans l'ensemble du secteur, et particulièrement au sein des équipes dirigeantes, on observe à la fois une inquiétude quant aux implications de sécurité liées à l'IA et un manque de capacité et de compétences pour y répondre.

À lui seul, ChatGPT a généré 410 millions de violations de politiques DLP dans les données de Zscaler. Dans le CISO AI Risk Report 2026 de Saviynt/Cybersecurity Insiders, 75 % des répondants ont découvert des outils de Shadow AI et 95 % doutent de leur capacité à détecter ou contenir les abus. Seul 1 % des entreprises dispose d'un budget dédié à la sécurité de l'IA (Pentera), et selon le Splunk 2026 CISO Report, bien qu'un consensus existe sur la capacité de l'IA à améliorer la productivité et à gérer les volumes, une majorité de RSSI s'inquiète des fuites de données, du Shadow AI et des hallucinations. Fait notable, Splunk observe que ces craintes sont plus élevées chez les RSSI des organisations les plus matures en matière d'IA, suggérant que ceux « qui sont plus avancés dans leur parcours d'IA générative commencent à percevoir certains compromis. »

L'IA est adoptée plus vite que les mécanismes de gouvernance prévus pour la contrôler, et le cadre réglementaire se durcit.

Ces compromis et ces craintes, en particulier concernant le Shadow AI, reposent sur des faits concrets. En avril 2026, Vercel a révélé une compromission déclenchée par l'utilisation de Context.ai par un employé, un outil tiers de productivité assisté par IA. L'employé s'était inscrit avec son compte Google Workspace professionnel et avait accordé des permissions « Autoriser tout » à l'application OAuth de Context.ai. Lorsque Context.ai a été compromis (apparemment via une infection Lumma Stealer sur l'appareil d'un employé de Context.ai) les attaquants ont hérité de ces jetons OAuth et ont pivoté vers les environnements internes de Vercel.

Le défi de gouvernance tient au fait que l'adoption de l'IA dépasse désormais les processus organisationnels conçus pour la gérer, alors même que l'environnement réglementaire se resserre et que l'écart entre la vitesse de déploiement et la maturité de gouvernance se creuse.
L'AI Act de l'UE exige que les systèmes à haut risque soient pleinement conformes en 2026, avec des sanctions pouvant atteindre 7 % du chiffre d'affaires mondial pour les contrevenants. Les exigences de transparence de l'état de Californie sont entrées en vigueur le 1er janvier 2026. Pourtant, selon Saviynt/Cybersecurity Insiders, bien que 71 % des grandes entreprises aient déployé des agents d'IA accédant aux systèmes métier critiques, seuls 16 % gouvernent cet accès de manière efficace.

Pour commencer à combler ce fossé, le trafic de prompts devrait être traité comme un journal de sécurité de première classe : la capture et l'analyse des prompts peuvent révéler des tentatives d'injection, des exfiltrations, des usages internes abusifs et des comportements inattendus d'agents. Cette approche peut débuter au niveau du proxy pour les modèles hébergés dans le cloud ou au niveau de la couche de harness pour les agents déployés localement, sans attendre une intégration complète au modèle. Mais les prompts ne constituent qu'une partie de la réponse. Les outils et les compétences deviennent une composante fondamentale de la surface d'attaque et doivent être traités comme tels.

Partie 3 :
L'impact de l'IA sur l'évolution des modèles opérationnels de défense

Triage et investigation

L'exécution de playbooks, la reconstruction de chronologies et le reporting en langage naturel sont des domaines où les modèles de raisonnement peuvent apporter une réelle valeur opérationnelle. Par exemple, l'infrastructure MDR de Kaspersky a traité environ 400 000 alertes en 2025, avec une logique de détection optimisée par l'IA filtrant les faux positifs avant la revue par un analyste ; seules 39 000 alertes ont été escaladées pour une investigation approfondie. De plus en plus, les modèles d'IA sont capables de mener ces investigations eux-mêmes. Prophet Security a mis en avant deux cas dans The Hacker News en mars 2026, où des modèles ont réalisé une analyse contextuelle et l'investigation d'incidents dont l'activité malveillante n'était pas immédiatement évidente.
Le premier cas concernait un utilisateur dormant doté d'une ancienne clé API qui s'est soudainement réactivée et le second consistait à détecter une intention malveillante dans un email de phishing qui aurait franchi la plupart (sinon l'ensemble) des contrôles de sécurité et de validation. Prophet Security souligne que dans les deux situations aucun point de données pris isolément n'était suspect. La menace n'est apparue qu'à travers une analyse contextuelle et multipoint — le type d'analyse que les enquêteurs humains réalisent lorsqu'ils disposent de la bande passante nécessaire. Or, cette bande passante fait souvent défaut, et c'est précisément là que les modèles de raisonnement peuvent contribuer, en offrant à chaque alerte le même niveau d'analyse.

Cependant, l'IA n'est pas une panacée et nécessite une ingénierie minutieuse. Les recherches sur l'overthinking montrent qu'au-delà d'un certain seuil, ajouter un raisonnement supplémentaire amène les modèles à inverser des réponses correctes. Les inversions négatives dépassent les positives après environ 7 000 jetons, atteignant un ratio de trois pour un à 12 000 jetons. Les implémentations des SOC devraient donc considérer l'effort de raisonnement, la limite de retry et les boucles d'appels d'outils comme des paramètres d'ingénierie ajustables.

Globalement, l'adoption reste inégale et les attentes doivent être calibrées. Seuls 40 % des RSSI utilisent actuellement l'IA générative dans leurs fonctions de sécurité, selon le rapport CISO de Splunk. L'IA agentique ne présente que 6 % de déploiement actif, bien que 39 % des entreprises l'étudient.

Il faut également reconnaître que les outils d'IA destinés aux équipes de défense peuvent eux-mêmes introduire leur propre surface de risque. Le rapport State of Pentesting de Cobalt, par exemple, a révélé que les applications d'IA et de LLM présentent des résultats à haut risque 2,7 fois plus souvent que les logiciels traditionnels, avec seulement 38 % des problèmes résolus lors des tests d'intrusion.

Le problème de la mémoire

La plupart des déploiements en production en sont encore à ce que certains chercheurs qualifient de « deuxième ou troisième génération 3 » de maturité agentique : des assistants augmentés par des outils, qui repartent de zéro à chaque session. Ils conservent leur apprentissage, mais ne peuvent pas se souvenir qu'un même schéma d'alerte apparu hier s'est révélé être bénin, ou que le blocage d'une plage d'adresses IP le mois dernier avait perturbé le VPN.

La distinction entre un analyste senior et un analyste junior ne tient pas nécessairement à leurs capacités de raisonnement, mais à ce qu'ils apportent à chaque événement : une mémoire épisodique des incidents passés, une connaissance sémantique de l'environnement et des intuitions procédurales sur ce qui fonctionne. Les assistants IA actuels sont plus juniors que seniors, et étudient chaque alerte comme si c'était la première fois. Les chercheurs étudient actuellement des architectures dont la mémoire est structurée comme un système d'exploitation (Memory-as-Operating-System), des services de mémoire externe conçus comme des intergiciels (Memory-as Middleware) et des approches basées sur des graphes de connaissances avec un raisonnement bi-temporel. Tant que ces approches ne seront pas matures, les systèmes IA resteront des assistants utiles, plutôt que des opérateurs autonomes.

La limite de détection

Les systèmes de détection comportementale opèrent sur des flux de télémétrie en temps réel marqués par un déséquilibre extrême des classes. Une plateforme traitant des billions (mille milliards) d'événements par jour agrège ces données en millions de détections dédupliquées, les priorise en milliers d'alertes de gravité élevée, et réalise des centaines d'investigations quotidiennes qui restent critiques après revue humaine. Le ratio de bout en bout approche un pour dix milliards. Les travaux fondateurs de Stefan Axelsson sur l'erreur de taux de base en matière de détection des intrusions ont formalisé la logique mathématique : même un détecteur doté d'une spécificité de 99,9 % génère des millions de faux positifs pour chaque véritable menace lorsque la probabilité préalable d'attaque est suffisamment faible.

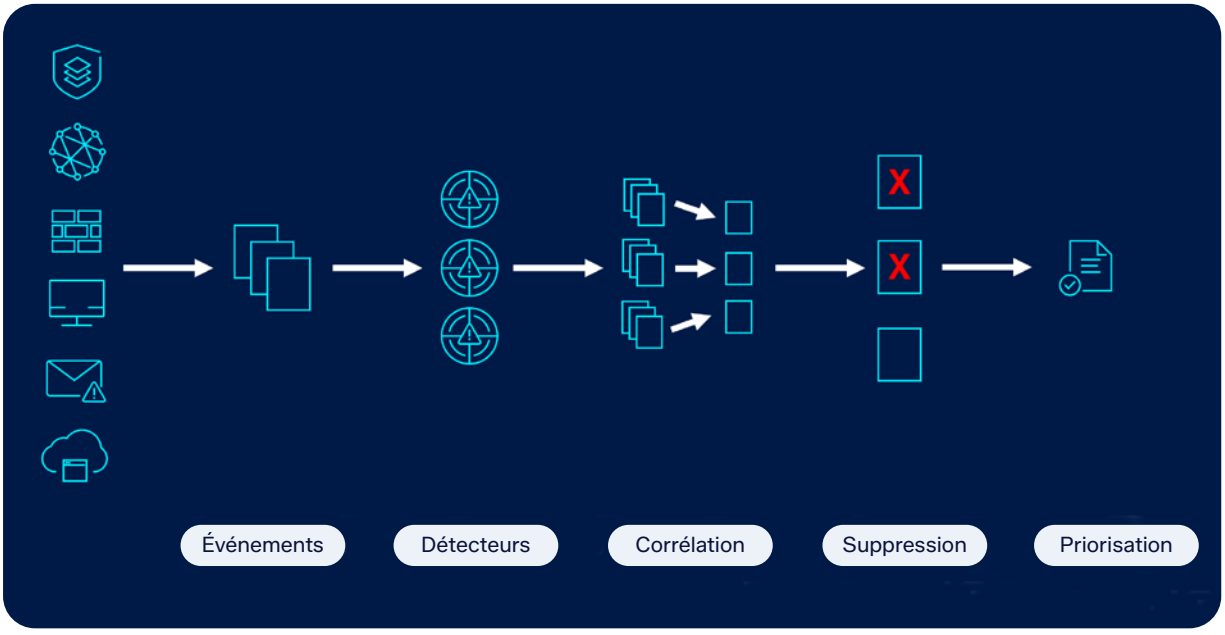


Figure 6 : Aperçu du pipeline en plusieurs étapes

Les travaux de recherche de Sophos présentés lors de la conférence NorthSec 2026 illustrent pourquoi la détection multicouche reste importante. Une fenêtre de deux semaines de télémétrie, représentant 11,8 billions d'événements, a été réduite via des détecteurs à la complexité croissante (correspondance d'indicateurs simples, règles de corrélation, modèles ML spécialisés, déduplication, corrélation et suppression) à 81 573 alertes de gravité élevée ou critique, soit moins de 50 par organisation en moyenne. Parmi ces alertes restantes, Sophos a identifié une attaque d'infostealer liée à la campagne TamperedChef.

Les modèles de raisonnement se situent au-dessus de cette couche. Ils génèrent des hypothèses d'investigation, récupèrent et synthétisent la télémétrie en résumés probants et coordonnent des actions de réponse limitées par des garde-fous explicites. Ils ne remplacent pas les modèles entraînés sur des données d'incidents propriétaires, des bases de référence spécifiques à l'environnement et des boucles de feedback continues auxquelles les systèmes de raisonnement généraux n'ont pas accès.

L'enquête AI Trust Maturity de McKinsey a quantifié la contrainte organisationnelle : les préoccupations en matière de sécurité constituent le principal obstacle à l'adoption à grande échelle de l'IA agentique pour près des deux tiers des répondants. L'inexactitude (74 %) et la cybersécurité (72 %) sont les principaux risques spécifiques. Seuls environ 30 % des organisations ont atteint le niveau 3 de maturité (« Des mesures sont prises pour développer l'ensemble des pratiques nécessaires en matière d'IA responsable ») ou au-delà. Les organisations disposant de responsabilités clairement définies en matière d'IA affichent un score moyen de maturité de 2,6, contre 1,8 pour celles dépourvues de gouvernance clairement définie.

La contrainte pesant sur la défense assistée par l'IA est la préparation organisationnelle. Les modèles étant suffisamment performants, la question est de savoir si les organisations peuvent les exploiter au niveau de confiance qu'exigent des opérations de sécurité autonomes.

Travaux de recherche de Sophos AI : Défendre la couche d'IA

Outre les travaux sur la détection multi-couche décrits plus haut, l'équipe Sophos IA a développé plusieurs autres techniques de défense inédites, chacune visant un mode de défaillance spécifique.

- **Le Untrusted Data Scanner (UDS)** inspecte le contenu externe pour détecter des tentatives d'injection de prompt avant qu'il n'atteigne un LLM ou un agent. Tout cas d'usage de LLM comporte un canal d'instruction (ce que l'application demande au modèle) et un canal de données (le contenu non fiable qu'il traite). L'injection de prompt dissimule des instructions dans ce canal de données. Le UDS est volontairement limité à ce problème précis, par opposition à un détecteur général de jailbreak. Les premiers résultats montrent un faible taux de faux positifs. Cela est important, car les données de sécurité regorgent de terminologie spécifique (rapports d'incident, analyses de malware, résultats de tests d'intrusion) et un détecteur qui « crierait au loup » à chaque rapport de menace serait inutilisable.
- Le **LLM Scoping** classe les requêtes selon qu'elles relèvent ou non de la fonction prévue d'une fonctionnalité, en rejetant celles qui sortent du champ d'application avant qu'elles n'atteignent le LLM. Parmi les méthodes évaluées, les modèles de scoping ont nettement surpassé les défenses basées sur le prompt système et ont réduit les taux de réussite des attaques à un niveau quasi nul sur plusieurs benchmarks publics. L'équipe a également appliqué le Multiround Automatic Red Teaming (MART) pour renforcer les modèles de scoping face à l'adaptation adverse.

- **CerBERTus**, un modèle BERT « à trois têtes », répond à la fragilité de la détection de jailbreak binaire. Un encodeur partagé alimente trois têtes de classification : la nocivité (principale), la catégorie d'objectif (ce que l'utilisateur cherche à faire) et le style de cadrage (la manière dont la requête est formulée). Les tâches auxiliaires agissent comme un biais inductif, incitant la représentation à séparer l'intention objective de son wrapper (enveloppe). L'équipe a établi un corpus structuré de prompts factoriels croisant des objectifs (cyberattaques, fraude, explosifs, synthèse de drogues, confidentialité, propagande extrémiste, etc.) avec des cadres adversariaux (jeu de rôle, fiction, urgence, prétexte académique, obfuscation) pour entraîner et éprouver cette séparation.
- **Détection des anomalies** : Dans l'étude présentée au Black Hat USA 2025, l'équipe a associé la détection des anomalies aux LLM et démontré que l'efficacité de la méthode réside dans la génération de diverses lignes de commande bénignes, plutôt qu'à l'identification de commandes malveillantes. Cette approche réduit fortement les taux de faux positifs dans les classifieurs de lignes de commande.
- Le **salage de LLM** (LLM salting), présenté lors du CAMLIS 2025, s'inspire du salage de mots de passe qui consiste à introduire des variations comportementales ciblées afin que des jailbreaks pré-calculés ne puissent pas être réutilisés sur des déploiements de LLM homogènes.

La tromperie contre les attaquants autonomes

Depuis vingt ans, la cyber-tromperie (Deception) n'a jamais été considérée comme un contrôle de premier plan. Mais face à des attaquants autonomes, l'équation change. Les agents d'IA ne se fatiguent pas et peuvent énumérer chaque chemin à la vitesse machine. Les environnements de tromperie, tels que ceux proposés dans le LLM Agent Honeypot de Palisade et le framework de MANTIS, peuvent saturer le contexte d'un agent avec des informations trompeuses, épuiser son budget d'inférence et entraîner des coûts économiques réels. HoneyTrap en est un exemple : il a démontré une augmentation de 149 % de la « consommation de ressources de l'attaquant » par rapport aux défenses classiques. Ce domaine en est encore à ses balbutiements, mais les conditions dans lesquelles la tromperie devient primaire sont précisément celles que crée l'essor de l'IA offensive.

Là où l'asymétrie de tempo se fait sentir

Un attaquant utilisant l'IA n'est soumis à aucun cycle d'acquisition, à aucun contrôle de conformité ni à aucun comité consultatif chargé des changements. Un acteur malveillant peut adopter un outil et l'opérationnaliser en quelques jours. La campagne assistée par IA visant les appareils FortiGate a compromis plus de 600 pare-feux dans 55 pays en cinq semaines, l'œuvre d'un acteur utilisant des outils d'IA commerciaux.

Côté défense, une détection se déclenche, l'investigation assistée par IA génère une recommandation de confinement en quelques minutes, mais le processus de remédiation du client nécessite un ticket de changement, une fenêtre de maintenance et un SLA de 48 heures. L'investigation s'est accélérée, mais la réponse n'a pas suivi. Cela semble avoir été le cas pour deux vulnérabilités exploitées très rapidement cette année : comme mentionné précédemment, le CVE-2026-10520 (Ivanti Sentry) a été exploité 24 heures après la publication du POC, et le CVE-2026-42208 (LiteLLM) 36 heures après.

Si le triage assisté par IA aide indéniablement les équipes de défense à gagner en tempo, le travail préparatoire sur les fondamentaux de la sécurité peut également contribuer à réduire cette asymétrie : diminuer la surface d'attaque, imposer l'authentification multifacteur et maintenir une télémétrie suffisamment complète pour reconstruire la chaîne des événements lorsque la prévention échoue.

Partie 4 : Recommandations pour les responsables de la cybersécurité

Les faits présentés dans de ce rapport permettent de prendre un nombre limité de décisions conditionnelles, en fonction de la situation actuelle d'une organisation.

Calibrer la tolérance au risque dans l'adoption de l'IA

Les faits exposés dans ce rapport créent une situation conflictuelle que les organisations doivent résoudre : l'adoption de l'IA introduit des risques tangibles, mais une adoption trop lente en introduit tout autant. Si l'on fait preuve d'une trop grande prudence alors que les attaquants gagnent en vitesse, on se retrouve à la traîne en matière de défense – et c'est précisément là, selon ce rapport, que réside le principal défi. Une prise de risque est donc inévitable, la question est de savoir comment prendre les « bons » risques.

Les faits exposés dans ce rapport créent une situation conflictuelle que les organisations doivent résoudre : l'adoption de l'IA introduit des risques tangibles, mais une adoption trop lente en introduit tout autant.

Le cadre de gestion des risques interne de Sophos distingue les « bons » risques des « mauvais » risques en s'appuyant sur deux axes : le potentiel de gain et la capacité à contenir les effets négatifs. Un « bon risque » a des répercussions tangibles, s'accompagne de mesures visant à limiter l'ampleur des dommages, d'une trajectoire réversible (pilote, restauration (rollback) ou repli (fallback)), de responsabilités clairement définies et de boucles de feedback bien établies. Un « mauvais risque » porte un potentiel de dommages illimités, enfreint des exigences non négociables (sécurité, confidentialité, juridique, conformité), ne dispose d'aucun plan de reprise, présente une responsabilité mal définie ou ne tient pas compte du discernement humain face à des changements qui affectent directement les clients.

Dans le cas de l'IA en particulier, ce cadre soulève, avant chaque déploiement, des questions pratiques : quel est le pire résultat envisageable ? Quelle est l'étendue potentielle des dégâts ? Peut-on revenir en arrière ? Qui en porte la responsabilité ? Les déploiements dont le risque est limité, par exemple un pilote en lecture seule d'un outil de triage agentique sur un jeu de données isolé, doivent être menés avec détermination. Les déploiements à haut risque, comme donner à un agent un accès en écriture à l'infrastructure de production, nécessitent une atténuation des risques ou une limitation du périmètre avant d'être mis en œuvre. Quant aux déploiements à faible impact mais à haut risque, ils doivent tout simplement être évités.

Les exigences non négociables s'appliquent quelles que soient les circonstances : confiance des clients, sécurité, confidentialité, conformité juridique et réglementaire, et stabilité opérationnelle. Le rôle du RSSI dans un environnement accéléré par l'IA est de veiller à ce que l'organisation prenne des risques intelligents, tout en évitant les risques inconsidérés.

Pour les organisations déployant des IA agentiques : réduire le rayon d'impact

Sophos a proposé sept contrôles aux organisations cherchant à déployer une IA agentique, réalisables dans un délai d'un à six mois :

1. **Le sandboxing des agents** définit un périmètre contrôlé autour du processus de l'agent. La plupart des harness intègrent une forme de sandboxing, mais celui-ci est généralement en « opt-in ». Dans la mesure du possible, privilégiez des environnements d'exécution distants ou Cloud plutôt que des solutions de sandboxing locales, afin de garantir une séparation plus forte des processus.
2. **L'isolement des identifiants** garantit que l'agent ne voit jamais les secrets directement : un processus distinct récupère les identifiants dans un coffre, les injecte dans l'appel API, puis ne renvoie que des réponses assainies, maintenant les identifiants longue durée hors des fenêtres de contexte des LLM.
3. **Les terminaux d'outils scellés** vont plus loin. L'agent appelle un outil par son nom, mais un broker détient l'identifiant, effectue l'appel API selon un schéma fixe et applique des listes d'autorisation des flux sortants propres à chaque outil. Le seul levier de l'agent est de choisir quel outil appeler et quels paramètres lui transmettre.
4. **La restriction des flux sortants et la surveillance du réseau** traitent l'agent compromis comme un canal d'exfiltration de données : détection de secrets dans le trafic sortant, seuils de volume sur les requêtes sortantes, et catégorisation web bloquant les requêtes vers des domaines non catégorisés ou récemment enregistrés.
5. **La couverture EDR** doit s'étendre au comportement des hôtes d'agents : conteneurs, machines virtuelles éphémères et postes développeurs non gérés constituent des angles morts fréquents.
6. **L'approbation subordonnée au contrôle humain** distingue le plan d'exécution autonome de l'agent d'un plan de contrôle géré par des humains : les opérations irréversibles requièrent un protocole d'autorisation cryptographique et non un simple bouton « OK » que l'agent serait capable de simuler.
7. **Les limites de propagation d'injection** répondent au risque croissant à mesure que les injections passent du périmètre de session à l'agent, puis à la propagation entre agents. Traitez les écritures en mémoire et les transferts entre agents comme des événements de sécurité.

Lorsque OpenClaw a gagné en popularité début 2026, plus de 30 000 instances étaient exposées sur Internet et les acteurs malveillants discutaient déjà de la manière d'utiliser ses « compétences » modulaires pour des campagnes de botnet. La Red Team de Sophos l'a testé directement : 23 enseignements actionnables, une réduction de la durée de la reconnaissance AD de trois jours à trois heures et un niveau de détail dans les audits impossible à atteindre manuellement dans un délai raisonnable. Mais l'équipe a passé plus de temps à établir des limites de sécurité qu'à effectuer le test. La conclusion est claire : l'IA agentique vaut la peine d'être déployée, à condition qu'un cadre de sécurité soit établi en premier.

Visibilité, identité et correctifs

Nous recommandons de commencer par obtenir une visibilité sur l'utilisation des services d'IA. L'exposition de l'IA mise en évidence dans les incidents Salesloft/Drift et Vercel en fait le premier levier à fort impact. Faites l'inventaire des outils d'IA présents dans votre environnement. Identifiez ceux qui détiennent des jetons OAuth, des clés API ou des comptes de service connectés aux systèmes d'entreprise. Déterminez quelles données transitent par eux et quelles autorisations ils détiennent. Vous ne pouvez pas gouverner ce que vous ne pouvez pas voir.

Le travail de visibilité doit aboutir à un registre de l'IA : chaque outil, agent, modèle, ainsi que les identifiants auxquels il peut accéder, la configuration de sandboxing et les besoins de connectivité attendus. Sans ce registre, des contrôles tels que l'isolement des identifiants, la restriction des flux sortants ou le sandboxing ne peuvent pas être appliqués de manière cohérente. Les contrôles d'identité doivent être appliqués avec la même rigueur que pour toute application SaaS : politiques d'accès conditionnel, MFA obligatoire pour les comptes de service IA, contrôles DLP sur les chemins des flux sortants utilisés par les outils IA, et rotation des identifiants des intégrations IA au même rythme que les autres accès privilégiés. Tant les conclusions de la X-Force d'IBM sur les données d'identifiants que l'attaque de la chaîne d'approvisionnement visant la Nx Console confirment que les identifiants d'accès des services d'IA font l'objet de ciblages actifs.

L'exploitation des vulnérabilités assistée par IA réduit à quelques heures seulement le délai entre la divulgation et l'exploitation.

Les organisations qui mettent encore des semaines à appliquer les correctifs s'exposent à des risques de plus en plus importants. Le délai de trois jours de la directive BOD 26-04 de la CISA constitue un seuil minimal, car l'exploitation des vulnérabilités assistée par l'IA réduit à quelques heures seulement le délai entre la divulgation et l'exploitation. Si une vulnérabilité critique exposée à Internet ne peut être corrigée en quelques jours, cet écart entraine un risque mesurable.

Évaluation basée sur les preuves

Dotez les outils d'IA utilisés par vos analystes de fonctionnalités de mesure. Mesurez l'effort de validation, le temps que les analystes consacrent à la vérification des résultats de l'IA. Mesurez l'efficacité des prompts, le nombre de tentatives avant d'obtenir des résultats utiles. Et mesurez le niveau de confiance, pour vérifier si les analystes font trop ou pas assez confiance au système. Créez des séries de tests basés sur les charges de travail réelles du SOC plutôt que sur des benchmarks génériques et gérez les versions de vos procédures afin qu'elles restent compatibles avec les mises à jour des modèles.

Sécurisez la chaîne d'approvisionnement de l'IA

L'IA introduit des risques liés à la chaîne d'approvisionnement qui vont au-delà des dépendances logicielles traditionnelles. Les poids des modèles, la provenance des données d'apprentissage, l'intégrité du serveur MCP et l'infrastructure d'inférence elle-même sont tous des limites de confiance. Les organisations déployant des modèles open-weight via des fournisseurs de services cloud managés doivent vérifier la version exacte du modèle, la région de fourniture du service, la durée de conservation des données, la politique relative à l'utilisation des données à des fins d'apprentissage, la surface de journalisation et les conditions contractuelles avant d'envoyer du code ou des données privées. Pour les modèles fournis via des terminaux SaaS hébergés en Chine, comme l'API de DeepSeek, un cadre plus strict s'impose. Ces modèles doivent être réservés aux tests de fonctionnalités avec des données publiques ou synthétiques, et non pour des référentiels propriétaires ou aux contextes d'ingénierie sensibles.

Conclusion

Le framework des acteurs STAC6994 regroupait des profils Cobalt Strike, des modules de contournement de l'EDR, des outils de collecte d'identifiants, des utilitaires de mouvement latéral et un canal C2 dissimulé derrière une infrastructure cloud légitime. Tous ces composants nous étaient familiers — sauf la manière dont ils avaient été produits. Les agents IA ont mis au point des techniques de contournement plus rapidement qu'un développeur humain n'aurait pu le faire, les ont testées en laboratoire dans un cadre structuré sur des produits spécifiques, ont géré les résultats via Git en différentes versions et ont transmis ces résultats à un opérateur chargé de déployer des ransomwares et de voler des données.

Que ce soit dans les dossiers traités par Sophos MDR et Sophos IR, les données de renseignement de la CTU, les analyses des SophosLabs, la recherche de Sophos AI ou le secteur de la sécurité en général, l'IA accélère les opérations courantes des deux côtés de l'équation de la sécurité. Elle permet aux attaquants d'avancer plus rapidement le long des chaînes d'attaque connues. Elle permet aux équipes de défense de trier et d'investiguer les incidents dans des délais extrêmement courts. Elle crée de nouvelles expositions au niveau des identités et de la gouvernance entourant l'adoption de l'IA en entreprise. Elle n'a pas produit de type d'intrusion fondamentalement nouveau que les architectures de détection actuelles ne seraient pas en mesure de gérer. Le AI Security Institute britannique a constaté, sur une période de 18 mois, une multiplication par six environ des capacités offensives autonomes. Chaque génération de modèle a permis de franchir des étapes dans des scénarios d'intrusion multi-étapes que les générations précédentes ne pouvaient pas accomplir. Cette trajectoire ne laisse aucune place à la complaisance.

Trois thèmes sont essentiels pour une mise en œuvre sûre de l'IA :

1. **La classification** doit être précise, car traiter toutes les menaces IA comme un bloc homogène il devient difficile de déterminer quelles mesures défensives sont nécessaires dans chaque cas.
2. **L'évaluation** doit être rigoureuse, qu'il s'agisse de routage de modèles, de suppression d'alertes ou de profondeur de raisonnement des agents, la réponse doit reposer sur des mesures.
3. **La transparence** doit être réelle, car la confiance naît au sein des entreprises et des organisations qui rendent leur travail public et tirent les leçons non seulement de leurs succès, mais aussi de leurs erreurs.

Ce que Sophos observera au cours des 12 prochains mois : si les agents autonomes s'écartent des scénarios d'attaque connus pour découvrir de nouvelles voies d'attaque à grande échelle, si l'économie souterraine développera des outils offensifs d'IA conçus comme des produits, abaissant ainsi le seuil d'accès de manière aussi radicale que le « ransomware-as-a-service » a abaissé celui de l'extorsion, si la gouvernance de l'IA en entreprise évoluera suffisamment rapidement pour combler la faille de sécurité liée aux identités générée par la vague actuelle d'adoption, et si le délai entre l'exploitation d'une vulnérabilité et la mise à disposition d'un correctif continuera de se raccourcir, imposant aux organisations qui mesurent encore leur temps de réaction en semaines à revoir leur approche.

Le tempo a changé. Les équipes de défense qui n'ajustent pas leur tempo se retrouveront plus à la traîne qu'elles ne le pensent.

Contributeurs

Nash Borges
SVP of
Engineering and AI

Adarsh Kyadige
Senior Manager
AI Research

Ross McKerchar
CISO

Rafe Pilling
Director of Threat
Intelligence
X-Ops Counter Threat Unit

Matt Stec
Director
Sophos AI

Ryan Westman
Senior Manager Threat
Research
X-Ops Insights

Matt Wixey
Senior Threat Researcher
X-Ops Insights

Notre équipe

Sophos X-Ops est une initiative de cybersécurité de pointe, réunissant plus de 1 000 experts issus de différents domaines de spécialisation au sein de Sophos : renseignements sur les menaces, intelligence artificielle, opérations MDR et opérations de sécurité internes.

Cette task force interfonctionnelle consolide les défenses des organisations face aux cyber menaces modernes, toujours plus rapides et sophistiquées.

En s'appuyant sur l'expertise croisée de ses équipes, Sophos X-Ops fournit une réponse multidimensionnelle aux attaques, assurant des capacités intégrées de protection, de détection et de réponse.

Cette approche collaborative et innovante fournit une réponse aux menaces d'un niveau inégalé, faisant de Sophos une référence en matière d'excellence et un leader de la cybersécurité.

Sophos X-Ops s'appuie sur l'expertise croisée de sa task force interfonctionnelle. La synergie entre les équipes de Sophos X-Ops nourrit un renseignement partagé, leur permettant d'intégrer rapidement de nouvelles informations, d'accélérer la détection et la réponse aux menaces, et de renforcer les capacités globales de protection pour les clients de Sophos.

SOPHOS

Pour discuter de vos besoins en matière de sécurité de l'IA et découvrir comment Sophos peut vous aider, consultez notre site Web ou contactez l'un de nos conseillers.

Sophos France
Tél. : 01 34 34 80 00
Email : info@sophos.fr

2026-07-20 FR (NP) CRE-5820