uniÿ

LIVRE BLANC

Cyber-AI

L'intelligence artificielle appliquée à la cybersécurité

Ce qui est réel, ce qui relève de l'effet de mode, et où investir

Un livre blanc **uniÿ**, studio d'architecture de systèmes agentiques

Lectorat : décideurs et clients · équipes d'ingénierie

Périmètre : organisations françaises, de l'ETI au groupe, secteurs régulés

Version de juillet 2026 · document périssable (voir Méthodologie)

Sommaire

Comment lire ce document

Ce livre blanc s'adresse à deux publics. Les passages de lecture stratégique (risques, gouvernance, retour sur investissement) intéressent en priorité les décideurs et les clients de nos prestations. Les passages d'approfondissement technique visent les équipes d'ingénierie. Aucune section n'est réservée : chacun peut tout lire, mais les encadrés signalent le niveau d'entrée.

Trois conventions structurent l'ensemble du texte :

> **FAIT, LECTURE ET RECOMMANDATION SONT DISTINGUÉS**
>
> Les affirmations factuelles, leurs interprétations et les recommandations qui en découlent sont séparées visuellement. Un fait est sourcé ; une lecture est une interprétation défendable mais discutable ; une recommandation engage un choix d'action.

Chaque constat porte un niveau de confiance. Nous utilisons quatre niveaux :

- `Établi` confirmé par plusieurs sources fiables et indépendantes, ou par une source primaire officielle.
- `Probable` appuyé par des sources sérieuses mais incomplètement recoupé, ou dépendant d'hypothèses.
- `Émergent` observé récemment, encore peu validé, susceptible d'évoluer vite.

- `Spéculatif` piste ou opinion non consolidée ; mentionnée pour vigilance, jamais traitée comme un fait.

Les sources sont tracées par famille. Quatre familles ont été croisées : **[Source A]** état de l'art académique ; **[Source B]** cas d'usage défensifs et produits ; **[Source C]** menaces offensives ; **[Source D]** gouvernance et cadre réglementaire. La veille terrain et les signaux faibles forment une cinquième famille, traitée à part au chapitre 7. La valeur de ce document tient à l'écart documenté entre ces sources : ce que la recherche promet, ce que le terrain observe, ce que la réglementation impose.

> GARDE-FOU ÉDITORIAL
>
> Ce document explique les mécanismes offensifs à un niveau strictement défensif et stratégique. Il ne fournit aucun mode opératoire d'attaque, exploit ou charge utile exploitable.

1. L'essentiel pour décider

Cette synthèse s'adresse en priorité aux décideurs. Elle résume ce qui est réel, ce qui relève de l'effet de mode, et où concentrer les investissements.

Ce qui est réel

L'IA défensive est aujourd'hui en production sur deux usages matures : le triage automatisé des alertes et les copilotes d'analyste qui rédigent, résument et permettent d'interroger les données en langage naturel. Les bénéfices crédibles sont des gains de productivité et de temps de réponse, pas une autonomie complète. Côté attaque, l'IA n'a pas encore créé de classe d'attaque radicalement nouvelle, mais elle abaisse massivement la barrière d'entrée et industrialise des modes opératoires existants : phishing irréprochable à grande échelle, deepfakes audio-vidéo

pour la fraude au virement, assistance au développement de code malveillant.

Ce qui relève de l'effet de mode

Le « SOC entièrement autonome » n'existe pas en production. L'analyste Gartner place les agents d'IA pour la sécurité au pic des attentes excessives et anticipe l'annulation de plus de 40 % des projets d'IA agentique d'ici fin 2027. Les chiffres les plus spectaculaires (précision de 98 %, réduction de 60 à 80 % des faux positifs) proviennent d'éditeurs ou d'études qu'ils ont commanditées, et doivent être relativisés. Le consensus de terrain est sans ambiguïté : l'IA reste un outil d'augmentation qui exige un humain dans la boucle, le risque dominant étant l'hallucination et la sur-automatisation.

Où investir

> RECOMMANDATION
>
> **1. Le triage et le copilote d'analyste,** pas le « SOC autonome ». C'est là que la valeur est démontrée et le risque maîtrisé.
>
> **2. Les fondamentaux qui neutralisent l'IA offensive :** authentification résistante au phishing (FIDO2/passkeys), validation multi-canal des virements, gestion des correctifs, détection comportementale (EDR/XDR), hygiène de la chaîne d'approvisionnement logicielle.
>
> **3. La gouvernance de la donnée et la souveraineté :** ne pas envoyer de données sensibles vers des LLM publics non maîtrisés ; privilégier des modèles auto-hébergés ou hébergés en cloud qualifié.
>
> **4. La conformité par anticipation :** l'AI Act et NIS2 ne sont pas encore pleinement contraignants en 2026, mais le sursis sert à constituer la preuve de conformité (cartographie, documentation, gouvernance).

> LECTURE

> Pour une organisation de taille intermédiaire, le retour sur investissement le plus sûr ne vient pas de l'IA la plus sophistiquée, mais de l'IA qui rend du temps à des équipes structurellement sous-dimensionnées. La sophistication offensive, elle, justifie surtout de renforcer des fondamentaux que beaucoup d'ETI/PME n'ont pas encore consolidés.

Pour situer votre propre exposition : le diagnostic en ligne cyberai.uniy.fr/diagnostic, moins de cinq minutes, sans création de compte.

2. Pourquoi la Cyber-AI, maintenant ?

Trois dynamiques convergent et expliquent que le sujet ne peut plus être différé.

Une bascule de la menace en 18 mois. En février 2024, Microsoft et OpenAI concluaient que les acteurs étatiques observés « greffaient » l'IA sur leurs modes opératoires sans en tirer de capacité offensive inédite[1]. En novembre 2025, Anthropic documentait ce qu'elle présente comme la première campagne d'espionnage largement orchestrée par une IA agentique[2]. L'IA est passée de conseiller à exécutant. C'est la trajectoire, plus que l'état instantané, qui doit guider la décision.

Une pression économique sur des équipes sous-dimensionnées. Les ETI/PME disposent rarement d'un SOC complet. L'argument économique de l'IA défensive n'est pas de remplacer des analystes inexistants, mais de redistribuer le temps rare des quelques personnes en place, en automatisant le triage et la rédaction.

Un cadre réglementaire qui se referme. L'AI Act européen, la directive NIS2 et le RGPD imposent progressivement des obligations de gestion du risque, de documentation et de notification. Les échéances les plus dures arrivent entre 2026 et 2028 (chapitre 6).

3. État de l’art : ce que la recherche établit, et ses angles morts

La recherche académique 2024-2026 envoie un message constant : l’écart entre les performances en laboratoire et la fiabilité en conditions réelles est structurel. Cette section intéresse en priorité les équipes d’ingénierie, mais sa conclusion est stratégique.

3.1 La détection par apprentissage : des scores en laboratoire à manier avec prudence

[Source A] **Établi** Les jeux de données de référence largement utilisés pour entraîner les systèmes de détection d’intrusion comportent des défauts documentés. Le corpus CIC-IDS2017, par exemple, contient des erreurs d’implémentation d’attaques, des erreurs d’outillage et des problèmes d’étiquetage qui gonflent artificiellement les performances rapportées[3].

[Source A] **Établi** Plus largement, une référence très citée recense les pièges méthodologiques récurrents (biais d’échantillonnage, fuite de données, évaluation en laboratoire seulement, métriques inadaptées) et conclut qu’une grande partie des résultats publiés ne reflète pas la performance opérationnelle[4].

> LECTURE
>
> Les scores de 99 % de détection que l’on rencontre dans les surveys et les plaquettes ne se transposent pas à un réseau d’entreprise réel. Pour une ETI/PME, la métrique qui compte n’est pas le taux de détection sur un jeu de données public, mais le taux de faux positifs supportable par une petite équipe.

3.2 Les LLM offensifs : capables, mais sous conditions strictes

[Source A] **Établi** Des agents fondés sur un grand modèle peuvent exploiter de façon autonome des vulnérabilités réelles déjà publiées : dans une étude, GPT-4 réussit 87 % de 15 vulnérabilités critiques *lorsque la description de la faille lui est fournie*, mais seulement 7 % sans cette description[5]. L'autonomie réelle est donc bien plus faible que le discours médiatique ne le suggère.

[Source A] **Probable** Sur les tâches de test d'intrusion complètes, les modèles échouent encore à enchaîner énumération, exploitation et élévation de privilèges sans assistance humaine[6].

[Source A] **Établi** Le code généré par des assistants d'IA comporte une proportion notable de faiblesses de sécurité : une étude empirique sur des dépôts GitHub relève des faiblesses dans environ 29,5 % des extraits Python et 24,2 % des extraits JavaScript examinés[7].

3.3 La sécurité des modèles eux-mêmes : deux résultats marquants

[Source A] **Établi** L'injection de prompt n'est pas résolue. Un benchmark systématique de cinq attaques et dix défenses conclut que les défenses existantes sont insuffisantes[8]. Sur une suite d'évaluation publique, tous les modèles testés cèdent à 26 à 41 % des tentatives d'injection[9].

[Source A] **Établi** L'empoisonnement est plus facile qu'on ne le croyait. Une étude conjointe Anthropic / UK AI Security Institute / Alan Turing Institute montre qu'environ 250 documents piégés suffisent à implanter une porte dérobée, et ce *indépendamment de la taille du modèle* (testé de 600 millions à 13 milliards de paramètres). Le nombre de documents requis ne croît pas avec la taille, ce qui renverse l'idée qu'il faudrait contrôler un pourcentage du corpus[10].

> **FAIT**
>
> Angle mort de la recherche : la quasi-totalité de ces résultats sont obtenus en laboratoire. Il manque cruellement d'études de terrain en

> conditions de SOC d'ETI/PME, et de jeux de données représentatifs d'un réseau de PME française.

4. L'IA côté défense : cas d'usage, bénéfices, pièges

Cette section cartographie l'usage réel de l'IA en défense, par niveau de maturité. Avertissement méthodologique : presque tous les chiffres de bénéfice disponibles proviennent d'éditeurs ou d'études qu'ils ont commanditées. Ils sont signalés comme tels et ne doivent jamais être lus comme des mesures indépendantes.

4.1 Matrice cas d'usage × maturité

Cas d'usage	Maturité	Bénéfice annoncé (et source)	Limite observée
Copilote d'analyste	Production	+22 % de rapidité, +7 % de précision ; −30 % de temps de réponse (Microsoft, étude interne)	Chiffres éditeur ; bénéfice surtout sur la rédaction et le résumé, pas sur la décision
Triage automatisé des alertes	Production	>98 % de précision de triage revendiquée (CrowdStrike, communiqué)	Communiqué marketing ; périmètre étroit, non reproductible hors écosystème
Réduction des faux positifs (UEBA, ML)	Pilote → Prod.	−60 à −80 % de faux positifs (synthèses secondaires)	Fourchettes larges, conditionnées au tuning et à la qualité des données
Threat intelligence	Pilote	Accélère la synthèse de renseignement (revue	Qualité variable ; risque d'hallucination

Cas d'usage	Maturité	Bénéfice annoncé (et source)	Limite observée
générative		académique)	de TTP/IOC ; vérification humaine indispensable
Orchestration agentique (SOAR + IA)	PoC → Pilote	Promesses d'automatisation bout-en-bout	Gartner : les promesses dépassent les preuves ; cas d'usage étroits

[Source B] **Probable** L'étude la plus solide du corpus éditeur est un essai randomisé contrôlé de Microsoft (janvier 2024) : analystes avec et sans copilote, tests chronométrés, protocole publié et co-signé par un chercheur externe. Elle donne +22 % de rapidité et +7 % de précision ; une étude de suivi sur 378 organisations donne −30 % de temps moyen de réponse[11]. Le retour sur investissement projeté par Forrester (de 372 k$ à 993 k$ sur trois ans) est une étude *commanditée par Microsoft* : à traiter comme une projection marketing.

4.2 Le contrepoint indépendant

[Source B] **Établi** Gartner place les agents d'IA pour la sécurité au pic des attentes excessives de son cycle 2025, juge que « les promesses dépassent les preuves d'amélioration mesurable et durable », et anticipe l'annulation de plus de 40 % des projets d'IA agentique d'ici fin 2027[12].

[Source B] **Probable** Le cadre indépendant MITRE ATLAS recense les attaques visant les systèmes d'IA eux-mêmes (empoisonnement, évasion, inversion de modèle). Déployer de l'IA en défense élargit donc aussi la surface d'attaque, un point que les plaquettes commerciales taisent.

> TENSION RECHERCHE / TERRAIN
>
> **Promesses éditeurs :** « 98 % de précision », « SOC autonome ».
>
> **Retour terrain :** une publication de mars 2026 décrit des déploiements peu profonds et des éditeurs qui attribuent les limites

> de leurs produits à la « psychologie de l'acheteur »[13].
>
> **Notre lecture :** traiter toute sortie d'IA comme un brouillon à valider, jamais comme une autorité. Un sondage Deloitte (2024) indique que 38 % de dirigeants ont pris une décision erronée sur la base d'une sortie d'IA hallucinée.

4.3 Pièges d'implémentation (pour l'ingénierie)

- **Hallucination en chaîne.** Une sortie hallucinée en threat intelligence (fausse TTP, faux indicateur de compromission) peut mésorienter toute une investigation.
- **Coût caché des workflows agentiques.** Les boucles de reprise (retry) génèrent typiquement 3 à 5 fois les jetons attendus ; une part importante de la dépense est gaspillée et rarement suivie (chapitre 7).
- **Faux négatif « confiant ».** Un agent qui documente parfaitement sa clôture d'alerte peut masquer une lacune de collecte plus difficile à rattraper qu'une clôture incertaine d'analyste humain.
- **Dépendance à la qualité des données.** Les gains annoncés sur les faux positifs supposent un référentiel comportemental propre et un tuning soigné : des compétences ML rares en PME.

5. L'IA côté attaque : panorama des menaces et contre-mesures

Cette section décrit, à un niveau strictement défensif, comment les attaquants emploient l'IA. Pour chaque menace : un niveau d'observation réelle (théorique, démontré en preuve de concept, ou observé en conditions réelles) et la contre-mesure associée. Aucun mode opératoire exploitable n'est fourni.

5.1 Taxonomie synthétique

Menace	Niveau d'observation	Contre-mesure prioritaire
Phishing / ingénierie sociale à l'échelle	Observé en conditions réelles	MFA résistante au phishing, filtrage avancé, sensibilisation aux signaux comportementaux
Deepfake / vishing (fraude au virement)	Observé en conditions réelles	Validation multi-canal et multi-personne des paiements, rappel sur numéro connu
Génération / assistance au malware	Observé (formes naissantes)	Détection comportementale (EDR/XDR), surveillance des appels sortants vers des API de LLM
Reconnaissance automatisée (acteurs étatiques)	Observé en conditions réelles	Réduction de l'exposition OSINT, gestion rigoureuse des correctifs
Agents offensifs autonomes	Observé (cas unique documenté)	Détection d'anomalies de volume/vitesse, moindre privilège, segmentation
Injection de prompt (directe/indirecte)	Démontré + sur produits déployés	Séparation instructions/données, moindre privilège des outils, validation humaine
Empoisonnement / porte dérobée	Démontré en preuve de concept	Traçabilité et curation des données, provenance des jeux de données
Slopsquatting (paquets hallucinés)	Démontré en recherche	Vérification des dépendances suggérées, allowlists, lockfiles, SBOM

5.2 Les trois menaces les plus concrètes pour une ETI/PME

[Source C] Établi **Phishing à grande échelle.** L'ENISA estime qu'au début 2025 plus de 80 % de l'ingénierie sociale observée est assistée par IA ; le phishing reste la porte d'entrée dominante des intrusions[14]. L'IA supprime les indices traditionnels (fautes, tournures maladroites) : la sensibilisation doit se recentrer sur les signaux comportementaux (urgence,

confidentialité, contournement de procédure) et non sur la qualité rédactionnelle.

[Source C] Établi **Deepfake et fraude au virement.** En 2024, un employé du bureau de Hong Kong de l'ingénieriste Arup a transféré l'équivalent d'environ 25 millions de dollars après une visioconférence où ses interlocuteurs étaient des deepfakes construits à partir de contenus publics[15]. La parade est organisationnelle : validation multi-canal de toute transaction sensible, codes convenus à l'avance, refus du caractère probant d'une simple visioconférence.

[Source C] Établi **Slopsquatting.** Une étude présentée à USENIX Security 2025 montre que près de 20 % du code généré par 16 modèles recommande au moins un paquet logiciel inexistant, et que 43 % de ces noms hallucinés réapparaissent de manière prédictible. Un attaquant peut enregistrer au préalable ces noms avec du code malveillant[16]. Pour une équipe qui adopte l'assistance au code, la vérification systématique des dépendances devient un réflexe de sécurité.

5.3 La menace de rupture : les agents offensifs autonomes

[Source C] Émergent Anthropic a documenté en novembre 2025 une campagne qu'elle attribue avec une confiance élevée à un acteur étatique, dans laquelle une IA agentique aurait exécuté l'essentiel des opérations contre une trentaine d'organisations, avec seulement quelques points de décision humains par campagne. L'entreprise note que le modèle hallucinait parfois des identifiants, ce qui a limité le taux de succès[17].

> LECTURE
>
> Cette divulgation est importante mais contestée (chapitre 7). Notre lecture : la donnée structurante n'est pas le pourcentage d'automatisation annoncé, mais la confirmation d'une trajectoire : l'agentivité offensive progresse. Les contre-mesures restent les fondamentaux du SOC : détection d'anomalies de volume et de

vitesse de requêtes, segmentation, moindre privilège pour limiter le mouvement latéral.

FAIT

Constat transversal : aucune des contre-mesures clés n'est spécifiquement « anti-IA ». L'IA augmente la fréquence et la crédibilité des attaques connues, pas leur nature. Les fondamentaux d'hygiène priment.

6. Gouvernance : ce que le droit exigera de vous, et quand

Une ETI/PME française qui emploie l'IA en cybersécurité évolue dans un empilement à deux étages : un étage transversal sur l'IA (AI Act) et un étage cyber (NIS2), sur un socle constant (RGPD). À cela s'ajoutent des référentiels volontaires structurants et un enjeu de souveraineté. Toutes les dates ci-dessous ont été vérifiées en juin 2026 ; certaines restent provisoires.

6.1 Obligations, échéances, impact

Texte / Norme	Obligation	Échéance	Impact ETI/PME
AI Act : pratiques interdites	Interdiction des IA à risque inacceptable	Applicable depuis le 2 févr. 2025	Vérifier qu'aucun outil RH/cyber ne fait du scoring comportemental interdit
AI Act : modèles d'usage général + gouvernance	Transparence, documentation amont	Applicable depuis le 2 août 2025	À exiger des fournisseurs de LLM ; le déployeur hérite

Texte / Norme	Obligation	Échéance	Impact ETI/PME
			d'obligations de transparence
AI Act : systèmes à haut risque	Conformité complète (risques, données, journalisation, supervision humaine)	Reporté du 2 août 2026 au 2 déc. 2027*	Sursis pour qualifier les usages « haut risque » et bâtir la documentation
AI Act : sanctions	Régime à trois niveaux	Selon échéances	Jusqu'à 35 M€ ou 7 % du CA mondial (pratiques interdites)
NIS2 (directive UE)	Gestion du risque, notification d'incident, responsabilité des dirigeants	Transposition due le 17 oct. 2024 ; France en retard	10 000 à 15 000 entités concernées en France, dont de nombreuses ETI/PME
NIS2 : loi française « Résilience »	Référentiel technique national (ReCyF)	Promulgation attendue été 2026*	S'auto-évaluer via le guichet MonEspaceNIS2 de l'ANSSI
RGPD + recommandations CNIL sur l'IA	Base légale, information, minimisation, droits	En vigueur ; avis CEPD déc. 2024	Tout LLM entraîné sur données clients ou journaux est concerné ; AIPD recommandée
ISO/IEC 42001:2023	Système de management de l'IA certifiable	Publiée fin 2023	Atout pour démontrer la maîtrise de l'IA ; s'intègre à une ISO 27001 existante
NIST AI RMF + profil GenAI	Cadre volontaire de gestion des risques IA	Profil GenAI publié le 26 juil. 2024	Méthode utile pour cadrer hallucination, injection, empoisonnement

Texte / Norme	Obligation	Échéance	Impact ETI/PME
ANSSI-PA-102	Recommandations de sécurité pour l'IA générative	Publié le 29 avr. 2024	Référence française opérationnelle pour sécuriser un déploiement LLM interne

** Échéances modifiées par l'accord politique « Digital Omnibus » du 7 mai 2026, qui reporte les obligations « haut risque » de l'Annexe III au 2 décembre 2027 et celles de l'Annexe I au 2 août 2028. Cet accord reste provisoire et doit encore être adopté et publié au Journal officiel de l'UE. De même, la date de promulgation de la loi française « Résilience » (transposant NIS2) est attendue à l'été 2026 mais n'était pas publiée au Journal officiel à la date de rédaction.*

6.2 Zones grises

- **Qualification « haut risque » d'une IA de cybersécurité.** Un EDR ou un SIEM augmenté n'est en principe pas concerné, mais une IA décisionnelle pilotant des contre-mesures sur une infrastructure critique pourrait l'être. La frontière reste à clarifier.
- **Statut « déployeur » contre « fournisseur ».** Une ETI qui affine (fine-tune) un modèle open-source peut basculer du statut de simple déployeur à celui de fournisseur, avec des obligations renforcées. Le seuil de modification « substantielle » n'est pas tranché.
- **Droit à l'effacement contre mémorisation des modèles.** La contradiction de fond entre le droit RGPD à l'effacement et la mémorisation par les LLM n'est pas résolue techniquement ; la CNIL le reconnaît.

6.3 Souveraineté et hébergement des modèles

`[Source D]` **Établi** Un LLM hébergé chez un fournisseur soumis au droit américain (même via un centre de données européen) expose les données à un accès extraterritorial (CLOUD Act). La qualification SecNumCloud de l'ANSSI exige explicitement l'immunité à ces lois.

Trois options s'offrent à une ETI/PME :

1. **LLM grand public en mode service :** rapide à déployer, mais souveraineté faible ; à proscrire pour les données sensibles (journaux, incidents, données clients).
2. **Modèles open-weight hébergés en cloud qualifié** SecNumCloud (acteurs français) : bon compromis souveraineté/fonctionnalités.
3. **Déploiement sur site ou isolé** pour les données les plus sensibles : maîtrise maximale, mais coût et compétences élevés.

> RECOMMANDATION
>
> L'ANSSI (recommandation ANSSI-PA-102) et la CNIL convergent : ne pas envoyer de données personnelles ou sensibles vers des LLM publics non maîtrisés. Pour l'usage cyber, privilégier des modèles auto-hébergés ou en cloud qualifié.

7. Signaux faibles et pistes émergentes

> POINT DE VIGILANCE
>
> Tout ce chapitre est prospectif. Les éléments ci-dessous proviennent de la veille terrain (forums, blogs de chercheurs, newsletters spécialisées). Ils sont **non consolidés** : signalés pour vigilance, jamais traités comme des faits. La crédibilité de chaque source est qualifiée.

7.1 Le désenchantement maîtrisé du terrain

Émergent **Le narratif « SOC autonome » s'est effondré** au profit d'un cadrage « l'IA redistribue le temps de l'analyste ». Les responsables sécurité sont décrits comme « légitimement sceptiques », se demandant

si l'IA traite vraiment les alertes ou n'est « qu'un moteur de règles déguisé »[18].

Émergent **L'« AI slop » sature l'open source.** Le mainteneur de curl a fermé son programme de bug bounty fin janvier 2026, environ 20 % des soumissions étant du contenu généré par IA techniquement crédible mais creux, pour seulement 5 % de vrais bugs[19]. Source crédible et indépendante, mais à ne pas sur-généraliser à tous les bug bounties.

Émergent **Le coût caché de l'agentique.** Plusieurs analyses FinOps estiment que les workflows agentiques génèrent 3 à 5 fois les jetons attendus, et que 40 à 60 % de la dépense serait gaspillée, alors que la moitié des organisations seulement suit son usage réel d'API. Chiffres à prendre comme ordres de grandeur (sources : éditeurs d'outils de maîtrise des coûts).

7.2 Un consensus d'experts indépendants : l'injection de prompt n'est pas résolue

Émergent Bruce Schneier et Simon Willison, parmi les voix les plus indépendantes du domaine, soutiennent que l'injection de prompt reste un problème non résolu, qui s'aggrave dès qu'on donne des outils à l'IA[20]. Sur l'écosystème des agents outillés (protocole MCP), un benchmark indique que les agents commerciaux les plus robustes échouent encore environ la moitié des scénarios d'injection via la sortie d'un outil. Ce sont des opinions argumentées et un domaine très récent, pas un théorème prouvé.

7.3 La controverse de la « première attaque orchestrée par IA »

> TENSION RECHERCHE / TERRAIN
>
> **Communication de laboratoire :** Anthropic présente une campagne « 80-90 % orchestrée par IA » (novembre 2025).

> **Réaction de chercheurs :** manque de détail technique, comportement (cadences de requêtes) jugé incompatible avec la furtivité d'un acteur sophistiqué, et conflit d'intérêt commercial relevé : les laboratoires qui divulguent vendent ensuite la défense[21].
>
> **Notre lecture :** à traiter comme une controverse, pas comme un fait établi. Une seule entreprise a la visibilité ; aucune validation indépendante publique à ce jour. La tendance de fond (montée de l'agentivité offensive) reste, elle, crédible.

8. Où investir : les décisions à arbitrer

8.1 Feuille de route pour l'ingénierie

1. **Commencer par le triage et le copilote.** Déployer l'IA là où la valeur est démontrée et le risque borné. Mesurer le bénéfice réel (temps gagné, faux positifs) sur un périmètre pilote avant toute généralisation.
2. **Imposer l'humain dans la boucle.** Toute sortie d'IA est un brouillon. Validation humaine obligatoire sur les actions à effet de bord et les décisions de réponse.
3. **Sécuriser les déploiements LLM internes.** Appliquer ANSSI-PA-102 : cloisonnement IA/SI, moindre privilège des outils accessibles à un agent, séparation instructions/données, journalisation, red teaming des garde-fous.
4. **Maîtriser la chaîne d'approvisionnement IA.** Vérifier les dépendances suggérées par l'IA (slopsquatting), tracer la provenance des jeux de données et des modèles (empoisonnement), tenir un inventaire (SBOM).
5. **Suivre les coûts.** Instrumenter la consommation de jetons des workflows agentiques dès le pilote pour éviter les dérives budgétaires.

8.2 Cadre de décision pour le client / décideur

1. **Renforcer les fondamentaux d'abord.** MFA résistante au phishing, validation multi-canal des virements, gestion des correctifs, EDR/XDR comportemental. Ce sont les contre-mesures qui neutralisent l'essentiel de l'IA offensive.
2. **Arbitrer la souveraineté selon la sensibilité des données.** LLM grand public pour le non sensible ; cloud qualifié ou hébergement isolé pour les journaux, incidents et données clients.
3. **Anticiper la conformité.** Cartographier les usages d'IA, qualifier le risque AI Act, s'auto-évaluer NIS2 via MonEspaceNIS2, documenter, pendant que les échéances dures ne sont pas encore tombées.
4. **Acheter avec scepticisme.** Exiger des éditeurs des preuves indépendantes, distinguer les chiffres commandités, prévoir un pilote mesuré. Se rappeler que plus de 40 % des projets d'IA agentique pourraient être annulés d'ici fin 2027.
5. **Investir dans les compétences.** La valeur de l'IA défensive dépend de la qualité des données et du tuning : sans compétence interne, le bénéfice annoncé ne se matérialise pas.

Situer votre exposition

Ce document dresse un état des lieux ; votre exposition, elle, dépend de vos décisions. Le diagnostic en ligne Cyber-AI la situe en moins de cinq minutes, sans création de compte, et rend une feuille de route priorisée : cyberai.uniy.fr/diagnostic.

Pour suivre ce risque dans la durée, le fil de veille publie des fiches sourcées, écrites pour ceux qui décident, accès validé à la main, sans email non sollicité : demande depuis cyberai.uniy.fr. Pour un échange sur votre situation : cyberai@uniy.fr.

Annexe A. Glossaire

Terme	Définition
Adversarial ML (apprentissage antagoniste)	Techniques visant à tromper un modèle (évasion à l'inférence, empoisonnement à l'entraînement, extraction).
Agent / IA agentique	Système d'IA capable d'enchaîner des actions de façon autonome via des outils, au-delà d'une simple réponse textuelle.
Deepfake	Contenu audio ou vidéo synthétique imitant une personne réelle.
EDR / XDR	Endpoint/Extended Detection and Response : détection et réponse comportementale sur les postes et au-delà.
Empoisonnement (data poisoning)	Insertion de données piégées dans un corpus d'entraînement pour induire un comportement déclenché.
Injection de prompt	Instructions malveillantes dissimulées dans les données qu'un modèle traite, le détournant de sa consigne.
LLM (grand modèle de langage)	Modèle d'IA générative entraîné sur de vastes corpus de texte.
MFA résistante au phishing	Authentification multifacteur non hameçonnable (FIDO2, passkeys).
SBOM	Software Bill of Materials : inventaire des composants logiciels d'un produit.
SIEM / SOAR	Collecte/corrélation d'événements (SIEM) et orchestration/automatisation de la réponse (SOAR).
Slopsquatting	Enregistrement malveillant de noms de paquets logiciels inexistants mais « hallucinés » de façon prédictible par l'IA.
SOC	Security Operations Center : centre de supervision de la sécurité.

Terme	Définition
UEBA	User and Entity Behavior Analytics : détection d'anomalies comportementales.

Annexe B. Bibliographie qualifiée

Références classées par famille. Niveau de confiance entre crochets. Liens vérifiés en juin 2026.

Famille A : État de l'art académique

- Engelen et al., « Troubleshooting the CICIDS2017 Dataset », IEEE S&P Workshops, 2021 · lien `Établi`
- Arp et al., « Dos and Don'ts of ML in Computer Security », USENIX Security, 2022 / ACM, 2024 · lien `Établi`
- Fang et al., « LLM Agents can Autonomously Exploit One-day Vulnerabilities », arXiv:2404.08144, 2024 · lien `Établi`
- « Towards Automated Penetration Testing (LLM Benchmark) », arXiv:2410.17141, 2024 · lien `Probable`
- Bhatt et al., « CyberSecEval 2 », arXiv:2404.13161, 2024 · lien `Établi`
- Fu et al., « Security Weaknesses of Copilot-Generated Code », arXiv:2310.02059 · lien `Établi`
- Liu et al., « Formalizing and Benchmarking Prompt Injection », USENIX Security, 2024 · lien `Établi`
- Souly, Rando, Carlini, Kirk et al., « Poisoning Attacks on LLMs… », arXiv:2510.07192, 2025 · lien `Établi`

Famille B : Cas d'usage défensifs

- Microsoft, étude de productivité Security Copilot (essai contrôlé randomisé, 2024 ; suivi 2025), source éditeur · lien `Probable`

- Gartner, communiqué du 25 juin 2025 ; Hype Cycle for Security Operations 2025 · lien `Établi`
- MITRE ATLAS (matrice des menaces sur les systèmes d'IA) · lien `Établi`
- Help Net Security, « AI SOC vendors are selling a future… », mars 2026 · lien `Émergent`

Famille C : Menaces offensives

- Microsoft, « Staying ahead of threat actors in the age of AI », 14 févr. 2024 · lien `Établi`
- ENISA, « Threat Landscape 2025 », oct. 2025 · lien `Établi`
- CNN Business, « Arup revealed as victim of $25 million deepfake scam », 16 mai 2024 · lien `Établi`
- Google Cloud / Mandiant, « GTIG AI Threat Tracker », nov. 2025 · lien `Probable`
- Anthropic, « Disrupting the first reported AI-orchestrated cyber espionage campaign », 13 nov. 2025 · lien `Émergent, contesté`
- Spracklen et al. (slopsquatting), USENIX Security, 2025 · lien `Établi`
- OWASP, « Top 10 for LLM Applications 2025 » · lien `Établi`

Famille D : Gouvernance et conformité (sources primaires)

- Règlement (UE) 2024/1689 (AI Act), EUR-Lex · lien `Établi`
- Conseil de l'UE, accord « Digital Omnibus » sur l'IA, 7 mai 2026, provisoire · lien `Probable`
- Directive (UE) 2022/2555 (NIS2) · lien `Établi`
- ANSSI, MonEspaceNIS2 et ReCyF (version de travail, 17 mars 2026) · lien `Probable`
- ANSSI-PA-102, « Recommandations de sécurité pour un système d'IA générative », 29 avr. 2024 · lien `Établi`
- NIST AI RMF 1.0 + profil GenAI NIST-AI-600-1, 26 juil. 2024 · lien `Établi`

- ISO/IEC 42001:2023 · lien `Établi`
- CNIL, recommandations sur l'IA ; avis CEPD, déc. 2024 · lien `Établi`

Famille E : Veille terrain (non consolidée)

- Schneier on Security, « Why AI Keeps Falling for Prompt Injection Attacks », janv. 2026 · lien `Émergent`
- Daniel Stenberg (curl) / The Register / Hacker News, sur l'« AI slop », 2025-2026 · lien `Émergent`
- tl;dr sec (Clint Gibler), éditions 2025 · lien `Émergent`
- « Evaluating LLM Generated Detection Rules », arXiv:2509.16749, 2025 · lien `Émergent`

Annexe C. Méthodologie de veille

Ce livre blanc a été produit par triangulation de quatre familles de sources (état de l'art académique, cas d'usage défensifs et produits, menaces offensives, cadre réglementaire), complétées par une veille terrain. Aucune affirmation structurante ne repose sur une source unique. Chaque constat porte une famille et un niveau de confiance. Les éléments non recoupés ont été écartés ou explicitement signalés.

Hiérarchie de confiance retenue, toutes choses égales par ailleurs : sources primaires officielles > publications évaluées par les pairs > rapports d'éditeurs (corrigés du biais commercial) > prépublications > forums et influenceurs. La fraîcheur d'un signal peut justifier de le remonter au rang de signal majeur, à condition de l'étiqueter comme non consolidé.

> **DOCUMENT PÉRISSABLE**
>
> Le sujet évolue vite. Les éléments les plus susceptibles de périmer en moins de douze mois sont : le calendrier de l'AI Act (accord « Digital

> Omnibus » encore provisoire), la transposition française de NIS2 (loi non promulguée à la date de rédaction), les capacités offensives des agents autonomes, et les performances revendiquées des produits. Une réactualisation trimestrielle de la veille réglementaire et de la veille terrain est recommandée.

Annexe D. Journal de vérification

Statut des affirmations clés au terme du contrôle final.

Affirmation clé	Statut	Note de vérification
AI Act : haut risque reporté au 2 déc. 2027	Confirmé (provisoire)	Accord politique du 7 mai 2026 recoupé ; texte non encore publié au JOUE
NIS2 : loi française attendue été 2026	Confirmé (provisoire)	Recoupé ; non promulguée à la date de rédaction
Empoisonnement : ~250 documents suffisent	Confirmé	Source primaire Anthropic/UK AISI/Turing recoupée
Fraude Arup ~25 M$ par deepfake	Confirmé	Couverture presse internationale concordante
Phishing >80 % assisté par IA	Cité avec réserve	Estimation ENISA ; méthodologie d'observation à expliciter
Microsoft Copilot : +22 % rapidité	Source éditeur	Étude commanditée ; protocole randomisé publié
Campagne « 80-90 % orchestrée par IA »	Contesté	Source unique (Anthropic) ; pas de validation indépendante ; controverse documentée
Gartner : >40 % projets agentiques annulés d'ici 2027	Confirmé	Communiqué Gartner du 25 juin 2025

Aucune URL n'a été inventée. Les chiffres non sourçables ont été écartés. Les contradictions entre recherche et terrain ont été conservées et signalées plutôt que tranchées arbitrairement.

---

1. Microsoft Security Blog, « Staying ahead of threat actors in the age of AI », 14 février 2024. ↩
2. Anthropic, « Disrupting the first reported AI-orchestrated cyber espionage campaign », 13 novembre 2025. Voir au chapitre 7 la controverse soulevée par cette divulgation. ↩
3. Engelen, Rimmer, Joosen, « Troubleshooting an Intrusion Detection Dataset: the CICIDS2017 Case Study », IEEE S&P Workshops (WTMC), 2021 ; confirmé par Lanvin et al., Springer, 2023. ↩
4. Arp et al., « Dos and Don'ts of Machine Learning in Computer Security », USENIX Security 2022, version étendue ACM 2024. ↩
5. Fang, Bindu, Gupta, Kang, « LLM Agents can Autonomously Exploit One-day Vulnerabilities », arXiv:2404.08144, 2024. ↩
6. « Towards Automated Penetration Testing: Introducing LLM Benchmark, Analysis, and Improvements », arXiv:2410.17141, 2024. ↩
7. Fu et al., « Security Weaknesses of Copilot-Generated Code in GitHub Projects », arXiv:2310.02059. Mise en contexte : CSET, « Cybersecurity Risks of AI-Generated Code », 2024. ↩
8. Liu et al., « Formalizing and Benchmarking Prompt Injection Attacks and Defenses », USENIX Security 2024. ↩
9. Bhatt et al., « CyberSecEval 2 », arXiv:2404.13161, 2024 (Meta PurpleLlama). ↩
10. Souly, Rando, Carlini, Kirk et al., « Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples », arXiv:2510.07192, 2025. ↩
11. Microsoft, étude de productivité Security Copilot (essai contrôlé randomisé, janvier 2024 ; étude de suivi, 2025). Source éditeur. ↩
12. Gartner, communiqué du 25 juin 2025 ; Hype Cycle for Security Operations 2025. ↩
13. Help Net Security, « AI SOC vendors are selling a future production deployments haven't reached yet », mars 2026. ↩
14. ENISA, « Threat Landscape 2025 », octobre 2025. ↩
15. CNN Business, « Arup revealed as victim of $25 million deepfake scam », 16 mai 2024. ↩
16. Spracklen et al., USENIX Security 2025 ; synthèses Cloud Security Alliance et Trend Micro, 2025-2026. ↩
17. Anthropic, « Disrupting the first reported AI-orchestrated cyber espionage campaign », 13 novembre 2025. Cette divulgation est contestée : voir chapitre 7. ↩
18. The Hacker News, séries « AI SOC » (oct. 2025 à janv. 2026) ; Detection at Scale (J. Naglieri), « 2025 Wrapped ». Sources à forte audience mais à intérêt commercial. ↩
19. Daniel Stenberg (curl), via The Register, BleepingComputer et fil Hacker News, juillet 2025 à janvier 2026. Source de crédibilité technique maximale, sans conflit commercial. ↩
20. Schneier on Security, « Why AI Keeps Falling for Prompt Injection Attacks », janvier 2026 ; Simon Willison, écrits 2024-2026. Commentateurs reconnus, sans conflit d'intérêt commercial notable. ↩

21. Critiques relayées par The Conversation (« Experts have questions »), IBM Think, PC Gamer, novembre-décembre 2025. ↩
