VEILLE IA & CYBERCRIMINALITÉ 11

MENACES ÉMERGENTES ET PRINCIPALES UTILISATIONS PAR LES CYBERCRIMINELS



▶ Ingénierie sociale et manipulation cognitive

Septembre confirme que l'IA permet aux fraudeurs d'aller au-delà de la simple génération de messages crédibles. En Italie, Reuters a révélé le 25 septembre une fraude ayant conduit Fideuram, filiale d'Intesa Sanpaolo, à transférer près de 95 millions d'euros. L'opération, initiée en février, combinait un faux message attribué à un dirigeant et un appel utilisant l'IA pour reproduire la voix d'un avocat connu de la victime. Environ 53 millions d'euros ont ensuite été récupérés, tandis qu'environ 36 millions restaient manquants selon Reuters.

Cette évolution est également illustrée par EvilTokens, plateforme criminelle démantelée par Microsoft. Elle utilisait l'IA pour analyser des messageries compromises, identifier les personnes de confiance, les responsables financiers et les circuits de paiement afin de préparer des scénarios d'usurpation ciblés. Microsoft indique que plus de 12 000 boîtes mail appartenant à plus de 10 000 organisations avaient été compromises.

Impact pour les victimes : L'IA permet de reproduire plus facilement des signes de confiance - voix, identité, relations professionnelles - et d'adapter rapidement une fraude au contexte réel de la victime.

Reuters, 25 septembre 2026
Microsoft Digital Crimes Unit, 22 septembre 2026

▶ Malwares et exploits générés ou pilotés par IA

L'utilisation d'agents IA commence à apparaître directement dans la conduite d'attaques. Le 22 septembre, Gambit Security a décrit une campagne visant des commerçants en ligne dans laquelle un acteur financièrement motivé utilisait plusieurs outils agentiques pour automatiser différentes étapes d'une intrusion. Entre le 10 et le 15 septembre, le laboratoire indique avoir observé plus d'une centaine de projets d'attaque et au moins 27 entreprises compromises à différents degrés.

Les données récupérées sur l'infrastructure analysée comprenaient également des informations liées à des cartes de paiement, dont plusieurs milliers émises en France. Le cas est particulièrement intéressant car l'IA n'est plus seulement utilisée pour rédiger du code ou préparer une attaque : elle contribue à enchaîner plusieurs tâches et à conduire simultanément des opérations contre un nombre important de cibles. L'ENISA observe plus largement dans son Threat Landscape 2026 les premiers signes de malwares intégrant des fonctions liées à l'IA, tout en soulignant que celle-ci sert encore principalement à accélérer des techniques déjà connues.

Impact pour les victimes : L'automatisation peut permettre à un même acteur de cibler davantage d'organisations simultanément et de réduire le temps nécessaire entre l'identification d'une cible et son exploitation.

Gambit Security Threat Intelligence, 22 septembre 2026
ENISA, Threat Landscape 2026, 22 septembre 2026

▶ Exfiltration et exploitation de données

L'IA modifie également la manière dont les données compromises peuvent être exploitées. Dans son rapport publié le 10 septembre, Anthropic décrit plusieurs opérations dans lesquelles des agents ont été utilisés pour rechercher des informations sensibles, identifier des secrets ou des accès privilégiés, préparer des extractions et analyser les données collectées.

Le rapport documente notamment un opérateur francophone ayant automatisé la recherche d'informations sensibles dans un très grand volume d'applications, ainsi que d'autres opérations où l'IA facilitait l'exploitation de données issues

d'environnements *SaaS*. Les activités décrites ont été observées entre décembre 2025 et août 2026 : elles ne constituent donc pas des incidents survenus en septembre, mais leur publication apporte ce mois-ci un éclairage important sur l'évolution des pratiques criminelles. L'enjeu n'est plus seulement de voler des données, mais de pouvoir les trier, rapprocher et exploiter rapidement afin d'identifier les informations présentant le plus de valeur.

Impact pour les victimes : Une fuite de données peut devenir plus dommageable lorsque l'IA permet d'identifier rapidement, parmi des milliers de documents, les personnes, projets, identifiants ou informations les plus sensibles.

Anthropic Threat Intelligence, 10 septembre 2026

▶ Attaques sur les modèles et infrastructures IA

Les infrastructures donnant accès aux modèles deviennent elles-mêmes une cible financière. Le 3 septembre, FortiGuard Labs a documenté un cas de *« LLMjacking »* dans lequel des identifiants cloud compromis ont été utilisés pour créer de nouveaux accès et consommer des modèles *via* Amazon Bedrock aux frais de la victime. Le principe rappelle le crypto-jacking : l'objectif n'est pas nécessairement de voler le modèle, mais de détourner une ressource informatique coûteuse pour son propre usage ou pour la revendre.

Un second signal concerne les composants reliant les agents IA aux applications. Le 28 septembre, un avis de sécurité GitHub a signalé une vulnérabilité de sévérité élevée dans le SDK Python du Model Context Protocol (MCP). Dans certaines conditions, un serveur MCP malveillant pouvait détourner des informations d'authentification utilisées pour accéder à des services tiers. L'avis documente une vulnérabilité confirmée, mais pas une campagne criminelle observée.

Impact pour les victimes : Les infrastructures IA introduisent de nouvelles formes de fraude à la consommation et augmentent la valeur des identités numériques donnant accès aux modèles, aux données et aux applications.

FortiGuard Labs, 3 septembre 2026
GitHub Security Advisory - MCP Python SDK, 28 septembre 2026

▶ Menaces hybrides et géopolitiques

L'IA contribue également à industrialiser certaines opérations d'influence. Le 28 septembre, NewsGuard a publié une analyse portant sur trois opérations (selon l'article en provenance de Russie) ayant diffusé de faux contenus attribués à des médias reconnus. L'étude recense 906 faux reportages attribués à 163 médias dans 21 pays, avec une exposition particulièrement importante de la France et de l'Allemagne. Huit des quinze médias les plus fréquemment imités étaient français.

L'intérêt de l'IA réside notamment dans la capacité à produire rapidement des contenus, des narrations et des variantes adaptées à différents publics. NewsGuard relève également un second phénomène : certains faux contenus peuvent ensuite être repris par des moteurs ou assistants IA, ce qui augmente leur capacité de diffusion au-delà des canaux utilisés initialement par les opérateurs. Cette tendance rejoint le constat d'ENISA, qui rapporte que 27 % des incidents de manipulation de l'information analysés par le Service Européen pour l'Action Extérieure (European External Action Service) en 2025 comportaient des techniques liées à l'IA, contre une proportion nettement plus faible l'année précédente.

Impact pour les victimes : L'IA facilite la production à grande échelle de contenus imitant des sources reconnues et peut rendre plus difficile l'identification de l'origine réelle d'une information.

NewsGuard, 28 septembre 2026
ENISA, Threat Landscape 2026, 22 septembre 2026

▶ Menaces émergentes et tendances à suivre

Le principal signal émergent de septembre concerne l'autonomie croissante des agents IA. Il ne s'agit pas encore d'une vague d'attaques criminelles autonomes observées dans la nature, mais de résultats expérimentaux qui montrent l'évolution rapide de leurs capacités.

Le 28 septembre, le UK AI Security Institute a publié une évaluation de GPT-6 Astra réalisée dans un environnement entièrement simulé et avec certaines protections cyber désactivées. Le modèle a conduit une attaque simulée contre une chaîne logicielle dans 29,2 % des trajectoires testées, avec notamment la création de fausses identités et des interactions avec des développeurs. OpenAI a parallèlement indiqué le 25 septembre qu'un modèle interne avait exploité une faiblesse de filtrage DNS pour communiquer avec un service externe depuis un environnement censé être isolé. L'incident s'est produit dans un environnement de recherche interne et ne correspond pas à une activité cybercriminelle.

Impact pour les victimes : Ces résultats ne décrivent pas une menace criminelle déjà généralisée. Ils montrent cependant que l'autonomie des agents et leur capacité à interagir avec des systèmes ou des personnes deviennent un risque à surveiller.

UK AI Security Institute, 28 septembre 2026
OpenAI Alignment, 25 septembre 2026

OPPORTUNITÉS ET OUTILS POTENTIELS POUR LES FSI

© COMCYBER-MI

INTERPOL - Operation Shams II

L'Operation Shams II, coordonnée par INTERPOL du 1er au 5 juin 2026, a mobilisé 28 officiers et experts de 11 pays pour exploiter l'IA dans l'analyse de contenus liés au terrorisme. L'IA a permis de trier et de dédupliquer plus de 108 000 images faciales, ramenées à 6 362 images uniques avant leur traitement par le système de reconnaissance faciale d'INTERPOL, avec vérification humaine des résultats. L'opération a conduit à l'identification de 126 combattants terroristes étrangers et a généré de nouvelles pistes d'enquête pour les pays participants. Ce cas d'usage illustre la montée en puissance d'un usage opérationnel de l'IA pour les enquêteurs spécialisés comme levier d'accélération du traitement et de l'exploitation de volumes massifs de données, difficilement accessibles par les seuls moyens humains.

INTERPOL, 18 septembre 2026

Europol - EnhanceOCR

EnhanceOCR utilise l'IA et le traitement d'image pour aider les enquêteurs à exploiter des plaques d'immatriculation difficilement lisibles sur des images ou vidéos dégradées. Présenté par Europol le 30 septembre, l'outil est expérimenté par la police néerlandaise et ses partenaires, avec vérification humaine des résultats proposés.

Europol, 30 septembre 2026

CERT-EU - Intelligent Ticket Routing

Le CERT-EU a déployé un système de routage automatisé des tickets cinq fois plus rapide que le triage manuel, pour orienter automatiquement les signalements de sécurité vers les équipes compétentes. Le système apprend en continu des décisions des analystes grâce à la récupération de tickets similaires et aux actions réalisées par les analystes. Par exemple, « lorsqu'un analyste déplace rapidement un ticket vers une file d'attente différente, cette correction est capturée comme un signal d'erreur. Lorsqu'une prédiction est laissée intacte pendant une période prolongée, malgré le fait que les analystes examinent régulièrement leurs files d'attente, le système interprète ce silence comme une validation implicite ». Cette solution permettrait aux FSI d'accélérer le délai de traitement initial d'un volume important d'informations.

CERT-EU, 22 septembre 2026

Commandement du ministère de l'Intérieur dans le cyberespace
#VeilleIA n° 11 - Octobre 2026

Directeur de la publication : GDI P. TOUAK
Responsable éditorial : LCL G. DECKMYN
Rédacteur : M. J. NEVICATO
Conception graphique et maquettage : Mme M. BARREAU

En attendant le prochain numéro de la veille IA & cybercriminalité, suivez notre actualité :

sur LinkedIn et sur X (ex-Twitter) : ***@ComCyberMI***