VEILLE IA & CYBERCRIMINALITÉ 10

MENACES ÉMERGENTES ET PRINCIPALES UTILISATIONS PAR LES CYBERCRIMINELS

© COMCYBER-MI

Ingénierie sociale et manipulation cognitive

L'IA industrialise les escroqueries et améliore leur crédibilité. INTERPOL estime dans son African Cyberthreat Assessment 2026 que l'IA facilite 55 % des cybercrimes signalés dans les pays étudiés. Les escroqueries restent la première forme de cybercriminalité observée, avec un recours croissant à l'IA pour produire des messages crédibles, automatiser l'ingénierie sociale ou générer des contenus synthétiques utilisés notamment dans la sextorsion. Europol souligne parallèlement que la fraude en ligne figure parmi les activités connaissant la croissance la plus rapide au sein du crime organisé. Pour les victimes, l'enjeu n'est donc pas seulement le deepfake spectaculaire : il réside surtout dans la capacité à cibler beaucoup plus de personnes avec des scénarios individualisés, cohérents et entretenus dans le temps.

Impact pour les victimes : davantage de citoyens peuvent être ciblés individuellement, avec un risque accru de pertes financières, d'usurpation d'identité et de revictimisation.

INTERPOL, 3 août 2026. INTERPOL report finds AI linked to more than half of cybercrime in Africa
Europol, 13 août 2026. How online fraud schemes are becoming a growing threat

Malwares et exploits générés ou pilotés par IA

L'IA devient surtout un accélérateur du travail des attaquants. Des opérations documentées durant l'été montrent son utilisation pour faciliter la reconnaissance, rechercher des vulnérabilités et produire ou adapter du code malveillant. Trend Micro rapporte notamment l'intégration croissante de l'IA générative dans les opérations de groupes étatiques. Il convient cependant de relativiser le phénomène des « malwares IA autonomes » : Unit 42 a étudié plus de 400 échantillons présentés comme intégrant de l'IA et environ 97 % n'avaient été observés que dans des environnements de recherche ou d'analyse. Le phénomène existe donc, mais reste encore limité dans les attaques réellement observées.

Impact pour les victimes : à court terme, le risque vient davantage de cybercriminels plus rapides et plus productifs grâce à l'IA que de logiciels malveillants entièrement autonomes.

Trend Micro, 29 juillet 2026. H1 2026 APT Activity Roundup
Palo Alto Networks Unit 42, août 2026. The State of AI-Enabled Malware

Exfiltration et exploitation de données

L'IA facilite l'exploitation des données après leur collecte. Une enquête publiée en août sur le groupe nord-coréen Kimsuky décrit l'utilisation de modèles locaux et de technologies permettant d'interroger et de synthétiser de grandes quantités de documents. L'intérêt pour un acteur malveillant est important : des données issues d'une intrusion peuvent être triées, rapprochées et analysées beaucoup plus rapidement. L'IA peut ainsi transformer une masse de données volées en informations directement exploitables : personnes à cibler, relations professionnelles, informations sensibles ou éléments utilisables dans une fraude ultérieure.

Impact pour les victimes : une fuite de données peut devenir plus dommageable lorsque l'IA permet de croiser automatiquement différentes informations et de reconstituer un profil détaillé d'une personne.

Reuters, 10 août 2026. North Korean hacking group builds AI tools for cyberattacks

Attaques sur les modèles et infrastructures IA

L'incident OpenAI-Hugging Face constitue l'un des signaux les plus importants de l'été sur les agents autonomes. Lors d'évaluations cyber, des modèles ont contourné les limitations de leur environnement, exploité des vulnérabilités et atteint des infrastructures réelles, notamment celles de Hugging Face. OpenAI précise que le modèle principal impliqué était un prototype interne, utilisé avec des protections volontairement réduites et non destiné au public. L'AI Security Institute britannique a par ailleurs rapporté 19 actions non autorisées sur Internet réalisées par des agents IA lors d'évaluations de

cybersécurité ; dans le cas le plus significatif, un agent a créé de fausses identités et tenté de convaincre un responsable humain d'intégrer du code malveillant dans un projet open source. Ces événements ne correspondent pas à l'utilisation normale des IA grand public, mais démontrent qu'un agent disposant d'outils, d'un accès Internet et d'une certaine autonomie peut sortir du périmètre prévu.

Impact pour les victimes : à terme, certaines opérations nécessitant aujourd'hui plusieurs interventions humaines pourraient être automatisées et conduites simultanément contre un nombre beaucoup plus important de cibles.

OpenAI, 26 août 2026. Hugging Face incident and the road ahead
UK AI Security Institute, 4 août 2026. Incident Report: unsanctioned agent behaviour during cyber testing

▶ Menaces hybrides et géopolitiques

Un cas concret d'utilisation d'agents IA dans une opération visant des institutions publiques a été confirmé à Taïwan. Le ministère du Numérique a indiqué le 13 août qu'une campagne d'origine étrangère, détectée dès juillet, avait combiné des techniques cyber conventionnelles et des agents IA pour attaquer des organismes gouvernementaux et des infrastructures critiques. Selon les autorités, ces agents permettent d'enchaîner rapidement différentes actions et rendent les campagnes moins coûteuses et plus faciles à industrialiser. Ce cas est important car il montre que l'IA n'est plus seulement utilisée pour préparer ou rédiger une attaque : elle commence à être intégrée directement dans son déroulement. Cette évolution rejoint les observations de Trend Micro sur l'intégration croissante de l'IA dans les opérations de groupes étatiques.

Impact pour les citoyens : la menace dépasse la compromission informatique individuelle ; elle peut toucher les administrations, les infrastructures essentielles et les données personnelles détenues par les services publics.

Ministry of Digital Affairs - Taiwan, 13 août 2026. Communiqué sur la campagne visant les administrations et infrastructures critiques
Trend Micro, 29 juillet 2026. H1 2026 APT Activity Roundup

▶ Menaces émergentes et tendances à suivre

La progression des capacités cyber des modèles ouverts constitue un point de vigilance. L'AI Security Institute britannique constate que l'écart entre certains modèles ouverts récents et les modèles propriétaires les plus performants se réduit sur certaines tâches cyber. L'enjeu criminel tient à leur disponibilité : un modèle ouvert peut être téléchargé, exécuté localement et modifié sans dépendre des contrôles exercés par son fournisseur. Cela pourrait progressivement rendre accessibles à des acteurs moins expérimentés certaines capacités aujourd'hui réservées à des spécialistes. Il s'agit toutefois d'une tendance à surveiller et non de la preuve d'une industrialisation criminelle déjà constatée.

Impact pour les victimes : la diffusion de capacités avancées pourrait abaisser progressivement le niveau de compétence nécessaire pour mener certaines attaques, sans que cette bascule soit encore démontrée à grande échelle.

UK AI Security Institute, 17 juillet 2026. How Far Behind the Frontier are Leading Open Weight Models on Cyber?

OPPORTUNITÉS ET OUTILS POTENTIELS POUR LES FSI

© COMCYBER-MI

▶ Évaluations cyber des modèles d'IA

Les évaluations menées notamment par le UK AI Security Institute permettent de mesurer ce que les modèles sont réellement capables d'accomplir dans des scénarios cyber. Elles peuvent être utiles aux FSI pour distinguer capacités démontrées, expérimentations et effets d'annonce, et pour anticiper l'évolution des capacités accessibles aux cybercriminels.

UK AI Security Institute, juillet-août 2026. Cyber evaluations

▶ Transparence européenne sur les contenus générés par IA

Les nouvelles règles européennes de transparence (AI Act) renforcent progressivement l'identification de certains contenus générés ou manipulés artificiellement. Elles présentent un intérêt pour les FSI dans le contexte des deepfakes, usurpations d'identité et manipulations informationnelles, en installant progressivement un cadre commun européen.

Commission européenne, 20 juillet 2026. Guidelines on AI transparency obligations

Commandement du ministère de l'Intérieur dans le cyberespace
#VeilleIA n° 10 - Septembre 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***