CYBERATTAQUES & IA

TENDANCES 2026

Les implications pour les organisations

CYBERATTAQUES & IA 2026

Tendances	Impact & Défense
Ingénierie sociale et hameçonnage basés sur l'IA Les pirates exploitent l'IA générative pour créer des e-mails d'hameçonnage très convaincants, des messages vocaux/vidéo deepfake et des interactions par chat qui imitent des personnes ou des marques de confiance.	Les filtres anti-hameçonnage traditionnels ont du mal à détecter ces messages, car ceux-ci ne présentent pas de fautes de grammaire ou de schémas linguistiques révélateurs. Analyse comportementale avancée, vérification d'identité pour les communications sensibles et formation des utilisateurs sur les *fake* possibles, comment les détecter et comment alerter.
Data Poisoning & Manipulation de Modèle AI Les pirates injectent des données malveillantes ou trompeuses dans les ensembles de données d'entraînement, ce qui corrompt l'intégrité des modèles d'IA.	Les modèles compromis peuvent prendre des décisions biaisées ou incorrectes, dégrader les performances, divulguer des informations sensibles. Utiliser des pipelines de données sécurisés, valider les ensembles de données et surveiller tout comportement anormal des modèles après leur déploiement.
Inversion et extraction de modèles Les pirates tentent de procéder à une ingénierie inverse ou de voler des modèles d'IA propriétaires en effectuant des requêtes répétées ou en exploitant des API exposées.	Perte de propriété intellectuelle, divulgation de données de formation sensibles (y compris des informations personnelles). Mettre en œuvre des techniques de limitation du débit, de « tatouage » numérique et de confidentialité différentielle ; sécuriser les endpoints des modèles grâce à l'authentification et à la surveillance.
Adversarial Machine Learning (AML) Les pirates créent des entrées subtiles et malveillantes, conçues pour tromper les modèles d'IA et les amener à mal classer les données (ex: système de reconnaissance d'images qui ne détecte pas une arme)	Sape la confiance dans les systèmes de sécurité basés sur l'IA, telle que la reconnaissance faciale ou la détection des logiciels malveillants. Formation robuste des modèles à l'aide d'exemples contradictoires, validation continue et modèles d'ensemble vérifiant les résultats entre eux.
Supply Chain AI et attaques par dépendance Les acteurs malveillants exploitent les composants d'IA tiers, les bibliothèques ou les modèles pré-entraînés que les organisations intègrent dans leurs systèmes.	Les portes dérobées ou vulnérabilités cachées dans les frameworks d'IA peuvent compromettre des environnements entiers. Effectuer une vérification approfondie des fournisseurs d'IA, tenir à jour les SBOM (nomenclatures logicielles) et surveiller tout comportement inhabituel dans les modules d'IA intégrés.

Sources

- Rapport IBM X-Force "Threat Intelligence Index, 2025".
- Akamai : "How AI Is Impacting the Fight Against Cybercrime", mai 2025
  Google/AP News, October 2025
- TrendsLab «State of AI Security 1H 2025»
- Blog Fortinet « CISO predictions 2026 », nov. 2025
- Tech Advisors, mai 2025 – AI Cyberattack Statistics
- Forbes "Cybersecurity 2026: 6 forecasts", nov 2025
- Blog CIS (Center for Internet Security) 2026-predictions
- Forecast Google Cloud "Cybersecurity Forecast 2026", nov 2025
- Forecast INE "2026 Cybersecurity Forecast", nov 2025<

LE HACKER 2026

Il attaque

- vos données
- vos modèles IA
- votre perception de la vérité
- vos émotions
- vos chaînes de dépendance

- Déformer, amplifier et propager le fake le plus fort et le plus vite possible
- Flouer vos repères culturels et informationnels

→ vos comportements et vos décisions, individuellement et collectivement

Pertes en domino

Crédibilité, Confiance

Cohésion, Engagement

Performance

Argent à court, moyen et long terme

Avoirs, Biens, Business… TOUT

Quand la vérité se fragmente, l'unité se désagrège.

LES IMPLICATIONS CONCRETES PAR TENDANCE

Et les questions à se poser impérativement
pour adapter sa communication de crise

La technology seule ne suffit plus à nous protéger.

Ingénierie sociale et hameçonnage basés sur l'IA

Le hacker d'aujourd'hui n'est plus un génie IT mais un **manager de campagne**, assisté par IA et par un éco-système de micro-prestataires sur des domaines spécifiques.

Le hacker travaille sur le scénario, le narratif. **Sa cible est l'humain** pour tromper sa perception.

- **Bypass MFA / Authentification**
- **Cloud Exploits** (Azure, AWS, GCP)
- **Firmware / ICS / OT**
- **Crypto-laundering**
- **Psychologie sociale / influence**

1 **Sommes-nous préparés à une attaque qui ne passe par aucune faille technique ?**

Exemple d'un faux document ou deepfake. Qui saura le détecter et quand ? Qui saura quoi faire ?

2 **Qui détient la vérité dans l'entreprise ?**

Si un deepfake ou faux rapport sort, qui est légitime pour dire que *"c'est faux"* ? **En combien de temps peut-il/elle répondre** ? Qui le croira ? Quels faux arguments peuvent être avancés ?

3 **Avec les outils IA actuels, qui dans l'entreprise** (ou hors d'elle) **peut produire un faux très crédible ?** Un ex-employé mécontent, un employé frustré, un stagiaire hacktiviste ? **Quelles compétences seraient nécessaires ?**

4 **Nos process de validation interne sont-ils IA-proof ?**

Si une IA interne publie automatiquement un contenu erroné/sensible, avons-nous un système de garde-fou ? Si un collaborateur publie par erreur une donnée sensible, avons-nous un droit de veto ?

5 **Combien de temps pour contrer le fake ?**

Si on publie un faux rapport, combien d'heures pour démontrer qu'il est truqué sans perdre la confiance ? Dans le monde IA, le temps de vérification devient le maillon faible. Les assaillants jouent sur cette inertie.

6 **Qui dans l'entreprise sait repérer les signaux faibles d'un contenu fabriqué** - image, son ou document **? Avons-nous un aperçu des manipulations possibles ?**

Au-delà des formations techniques et sensibilisation, qui sait gérer une attaque de crédibilité orchestrée ?

7 **Avons-nous une stratégie d'alliés crédibles pour contrer la désinformation ? Quels tiers de confiance** (journalistes, partenaires, experts reconnus) **peuvent relayer un démenti rapidement et efficacement ?**

La guerre d'influence se gagne avant la crise par la construction d'un réseau de confiance externe.

8 **Nos visuels, vidéos, voix officielles sont-elles horo-datées, signées, traçables ?** Les contenus officiels authen-tifiés doivent être placés dans un coffre-fort dédié pour prouver l'authenticité d'un contenu officiel en cas de deepfake.

Data Poisoning & Manipulation de Modèle AI

Le hacker cherche à **altérer le fonctionnement ou la crédibilité de l'IA** utilisée par l'entreprise pour prendre des décisions cruciales (prix, RH, com interne, prévisions, clients…).

Sa cible est l'entraînement ou l'intégrité du modèle pour corrompre la "mémoire" ou la logique interne de l'IA.

C'est une attaque contre la formation ou la logique interne de l'IA qui touche la stratégie de l'entreprise.

1. **Pouvons-nous prouver rapidement qu'un modèle a été corrompu, d'en identifier l'origine** (interne, externe, fournisseur) **avec quel l'impact sur notre responsabilité légale et notre assurance en cas de dommage ?** Les lois (AI Act, DORA, etc.) à venir imposent de tracer et auditer les données utilisées.

2. **Veillons-nous aux data entrant dans nos modèles et aux process de réentraînement automatisé ? Avons-nous un mécanisme détectant à temps une contamination du dataset, une anomalie du modèle ?** La plupart des entreprises surveillent leurs serveurs, pas leurs modèles.

3. **Nos équipes savent-elles que l'IA peut être manipulée sans être détectée et que son comportement peut changer sans alerte technique ? Savent-elles identifier les signaux faibles d'un modèle corrompu ?** La vigilance doit faire partie intégrante de la culture, savoir vérifier ce que dit l'IA avant d'agir en conséquence.

4. **Si notre IA de pricing, de production ou de communication est compromise, combien de temps avant de s'en rendre compte ? Avons-nous un plan B sans IA ? Savons-nous geler un modèle en crise ?**
Beaucoup d'entreprises dépendent de systèmes IA sans savoir comment reprendre la main.

5. **Pouvons-nous expliquer à un tiers une décision erronée issue d'un modèle contaminé ? Avons-nous la capacité d'auditer et retracer les données d'entraînement pour démontrer une manipulation** en cas d'enquête ou de crise publique **?** Le risque est réputationnel et moral avec perte de confiance.

6. **Qui contrôle nos sources de données externes ?**
Les poisoning modernes viennent souvent d'un fournisseur, d'un open dataset, d'un modèle pré-entraîné ou d'un agent IA externe. Aucune entreprise ne maîtrise toute sa chaîne.

Inversion et extraction de modèles

Le hacker **vise ce que le modèle sait**, ce qu'il a appris dans les simulations.

Il **attaque la mémoire du modèle** pour reconstruire des informations sensibles.

Il chercher à **voler la propriété intellectuelle invisible** en s'attaquant au modèle lui-même (logique, structure, comportement, paramètres, etc).

Objectifs : revente de données, sabotage de la confiance dans le modèle, rançon, espionnage industriel.

C'est une attaque sur les opérations business, qui touche la confidentialité et la propriété de l'entreprise.

1 **Quel modèle d'IA, agent ou outil interne constitue une véritable propriété intellectuelle, et lesquels peuvent être interrogés depuis l'extérieur** (API, assistants, bots, agents) **même indirectement ?** Inventorier/qualifier ses outils d'IA est vital pour identifier les plus exposés ou critiques, à protéger sur le plan technique et juridique.

2 **Que signifierait la perte du modèle pour votre business et votre réputation à court, moyen et long terme ?** L'analyse des impacts permet d'identifier clairement les risques et de les hiérarchiser afin de mettre en œuvre un plan de continuité et un pare-feu réputationnel appropriés.

3 **Comment gérer une crise, où un modèle volé ou cloné est utilisé pour produire des contenus qui semblent provenir de nous ?** Une gouvernance de vérité et une stratégie de preuve deviennent cruciales pour maintenir la confiance des interlocuteurs internes externes.

4 **Avons-nous des preuves techniques pour démontrer qu'un modèle volé et modifié n'est pas le nôtre (signature, hash, versioning sécurisé, journaux, empreintes cryptographiques) ?** Tout ce qui peut permettre de faire de solides déclarations publiques, et crédibles.

5 **Si un modèle volé divulgue des données sensibles ou si une inversion de modèle expose nos données d'entraînement, sommes-nous capables d'en identifier l'origine, d'en assumer la responsabilité et de répondre aux obligations réglementaires (AI Act, RGPD) ?**

6 **Nos endpoints IA exposés sont-ils protégés et monitorés pour bloquer une extraction de modèle ?** Contrôlons-nous le volume, les patterns et la nature des requêtes adressées à nos modèles (API, agents, bots) ? Appliquons-nous du throttling*, une segmentation et un monitoring comportemental pour détecter toute extraction ?

*throttling : mécanisme qui limite ou ralentit le volume des requêtes adressées à un système pour empêcher les abus, la surcharge ou les extractions malveillantes.

Adversarial Machine Learning (AML)

Le hacker **devient manipulateur d'IA** pour fausser vos modèles qui influencent les décisions.

Il fausse les capteurs, déforme les perceptions pour induire un mauvais choix.

Sa cible est le modèle IA lui-même pour tromper la perception de l'IA.

C'est une attaque sur les opérations stratégiques business de l'entreprise.

1 **Quelles décisions critiques sont prises par quel modèle IA ? Quels modèles IA prennent des décisions automatiques sans revalidation humaine ?** Avons-nous cartographié les modèles en autonomie décisionnelle ? Si l'IA décide seule = danger. Si l'humain revalide = l'attaque échoue, crée seulement une perturbation.

2 **Par où ou par quoi un contenu piégé** (image, son, texte, fichier, capteur, API) **peut tromper notre IA, et pouvons-nous rejouer la décision pour comprendre ce qui a été manipulé ?** Une attaque AML passe par un élément à peine détectable. Tracer l'incident pour reconstruire après coup la séquence de décision aide à corriger.

3 **Notre modèle est-il entraîné pour reconnaître qu'il est manipulé ? Avons-nous un protocole de tests adversariaux ?** Très peu d'organisations testent leur modèle avec des perturbations adversariales, vérifient sa robustesse, effectuent du red-teaming IA.

4 **Avons-nous une "zone tampon" pour confirmer une décision douteuse ? Qui détecte l'étrangeté ? Y a-t-il un seuil d'alerte ? Qui est habilité à désactiver un modèle compromis ?** Une attaque AML produit une décision surprenante, incohérente, et statistiquement improbable. Existe-t-il un protocole de "*stop-the-line IA*" ?

5 **Nos équipes savent-elles reconnaître une réponse aberrante… ou vont-elles l'utiliser sans la questionner ? La culture interne inclut-elle la vigilance anti-IA ?** C'est un angle mort fréquent, où les équipes ont tendance à faire trop confiance au modèle.

6 **Savons-nous communiquer, à l'interne et l'externe une erreur AML sans paniquer l'opinion ? Avons-nous un narratif prêt pour une "erreur IA" ?**

C'est une crise à part entière, qu'il s'agisse d'une erreur de diagnostic, de pricing ou une recommandation biaisée.

Supply Chain IA/Cloud et attaques par dépendance

Le hacker **exploite la structure en étoile d'un fournisseur central et profite de la lenteur de la détection** et de la dilution des responsabilités pour multiplier le nombre de victimes avant que l'alerte ne soit donnée.

La solution est industrialisée et déployée après la compromission.

La **cible, ce sont les couches invisibles de l'IA** (infrastructures cloud multi-locataires, appareils connectés mal sécurisés, pipelines de données, IA autonomes non supervisées).

C'est le principe du *"One-to-many attack" et "Single point of compromise" avec propagation massive*

1 **Où se trouvent nos données ? Combien d'IA, agents IA les manipulent ? Comment est validée la fiabilité des data qui alimentent nos IA ? Avons-nous cartographié les modèles en autonomie décisionnelle et les pipelines de data ?** Il faut identifier la surface invisible, prioritiser les composants et zones pour prévenir les contaminations.

2 **Avons-nous une visibilité inter-systèmes** (IoT, IA, cloud) **pour repérer en temps utile les comportements anormaux ? Nos agents IA ont-ils des garde-fous ou peuvent-ils agir sans supervision humaine ?** Il faut passer de la surveillance locale à l'observation systémique, et évaluer le risque d'autonomie incontrôlée.

3 **Qui contrôle la sécurité dans notre chaîne cloud** (multi-locataire ou non ?) **? Le fournisseur, nous, ou personne ? Si un maillon de la chaîne cloud tombe, en combien de temps perdons-nous la continuité métier ?** Il faut clarifier les responsabilités contractuelles et anticiper la propagation d'une panne ou d'une manipulation.

4 **Nos prestataires sont-ils soumis à un calendrier de patchs aussi strict que le nôtre ? Comment vérifier que ceux annoncés sont bien appliqués sur toute la chaîne de dépendance ?** Anticiper la propagation d'une panne ou d'une manipulation et vérifier la chaîne de dépendances pour bien anticiper la communication.

5 **Savons-nous expliquer clairement la panne ou la manipulation venant de la supply chain IA ? Notre communication est-elle alignée avec celle des fournisseurs cloud/IA, en cas d'incident ? Qui parle si l'un deux se tait ou contredit notre version ?** Toute cette communication doit être préparée à l'avance.

6 **Y a-t-il une procédure pour informer les clients, partenaires et régulateurs, même si nous ne maîtrisons pas toute la chaîne de dépendances ? Avons-nous un narratif prêt contre "vous auriez dû mieux contrôler votre IA" ?** Dans une attaque supply chain IA, l'entreprise reste perçue comme responsable, même si la cause est externe.

LE NOUVEAU RISQUE EST INVISIBLE

Avec la pression interne comme surface d'attaque

- Sous pression, les équipes deviennent plus manipulables, moins vigilantes.
- L'IA amplifie exactement les signaux qui déstabilisent : confusion, surcharge, doute.

Les hackers exploitent ce que votre management fragilise

SIX LEVIERS IMMÉDIATS

Les fondations indispensables d'une défense moderne

Gouvernance de la vérité et stratégie de la preuve

Définir comment on établit ce qui est vrai, ce qui est douteux, qui tranche et comment. Protéger & authentifier les contenus officiels sensibles.

Circuit interne d'alertes & de veille critiques (spécial PME)

Canal simple pour remonter incidents, signaux faibles, incohérences IA et éléments repérés en ligne. Agréger, filtrer, diffuser l'info utile, même sans équipe cyber.

Résilience élargie à l'humain, au culturel et aux réflexes de décision

Approche psychologique et systémique de la résilience pour cultiver vigilance, cohésion et réflexes efficaces en situation d'incertitude, de manipulation.

Maîtrise de sa communication sous pression

Entraîner les dirigeants et porte-parole à gérer doute, confusion ou désinformation. Une parole structurée dès les premières minutes change tout.

Tests et Simulations en conditions réelles

Scénarios IA, faux messages, deepfakes, vishing, signaux contradictoires, etc. On ne teste pas son plan le jour de la crise.

Esprit critique & réflexes anti-manipulation

Former les équipes à repérer les signaux cognitifs faux, réagir sans panique, adopter les réflexes de protection et signaler factuellement les anomalies.

Un leadership préparé, sans budget cyber massif.

QUI EST OZ'N'GO

Votre partenaire en Communication de Crise & Résilience Cyber.

Maryse Rebillot

Fondatrice
Résilience & Communication de Crise

- 20+ ans en marketing-communication dans le High Tech et la Cybersécurité
- Experte en gestion de crise & Communication sensible
- International Executive Marketing Master, INSEAD

linkedin.com/in/maryse-rebillot

Réseau d'Experts

(mobilisables selon les missions, en fonction des besoins)

Cybersécurité

Cyber défense, White Hacking, Gouvernance Cyber, Gouvernance des données, Légal & Assurance

IA & Technologies

IA Générative, LLM, IA Vertical, Autres IA

Humain & Communication

Psychologie, Médias & Journalistes, Sociologie

Oz'n'Go, le guide des dirigeants pour maîtriser les crises que l'IA crée, amplifie ou manipule.

À L'ÈRE DE L'IA,
LA SEULE DÉFENSE DURABLE

c'est la cohésion humaine.

Maryse Rebillot

contact@ozngo.com

+41 77 533 77 67

www.ozngo.com