CYBERATTACKS & AI

2026 TRENDS

Implications for organisations

CYBERATTACKS & AI 2026

Oz'n'go

Trends	Impact & Defense
AI-Powered Social Engineering & Phishing Attackers are leveraging generative AI to create highly convincing phishing emails, deepfake voice/video messages, and chat interactions that mimic trusted individuals or brands.	🎯 Traditional anti-phishing filters struggle because these messages lack telltale grammar or language patterns. 🛡 Advanced behavioral analytics, identity verification for sensitive communications, and user awareness training augmented by AI-based detection.
Data Poisoning and Model Manipulation Attackers attempt to reverse-engineer or steal proprietary AI models through repeated queries or exploitation of exposed APIs.	🎯 Loss of intellectual property, exposure of sensitive training data (including personal information). 🛡 Implement rate limiting, watermarking, and differential privacy techniques; secure model endpoints with authentication and monitoring.
Model Inversion and Extraction Attackers attempt to reverse-engineer or steal proprietary AI models through repeated queries or exploitation of exposed APIs.	🎯 Loss of intellectual property, exposure of sensitive training data (including personal information). 🛡 Implement rate limiting, watermarking, and differential privacy techniques; secure model endpoints with authentication and monitoring.
Adversarial Machine Learning (AML) Attackers craft subtle, malicious inputs designed to fool AI models into misclassifying data (e.g., an image recognition system failing to detect a weapon).	🎯 Undermines trust in AI-driven security systems like facial recognition or malware detection. 🛡 Robust model training with adversarial examples, continuous validation, and ensemble models that cross-verify outputs.
AI Supply Chain and Dependency Attacks Threat actors exploit third-party AI components, libraries, or pre-trained models that organizations integrate into their systems.	🎯 Hidden backdoors or vulnerabilities in AI frameworks can compromise entire environments. 🛡 Conduct thorough vetting of AI vendors, maintain SBOMs (Software Bills of Materials), and monitor for unusual behavior in integrated AI modules.

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<0

08/12/2025

THE 2026 HACKER

He attacks

- your data
- your AI models
- your trugth perception
- your emotions
- Your dépendance chain

- 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

Domino effect Losses

Credibility, Trust

Cohesion, Engagement

Performance

Money in short, medium, and long term

Assets, business... EVERYTHING

When truth fragments, unity disintegrates.

IMPLICATIONS
PER TREND

And questions you must ask yourself in order to adapt your crisis communication

Technology alone is no longer enough to protect us.

AI-Powered Social Engineering & Phishing

Today's hacker is no longer an IT genius but a **campaign manager,** assisted by AI and an ecosystem of micro-service providers in specific fields.

The hacker works on the scenario, the narrative. The target is humans, to deceive their perception.

- **Bypass MFA / Authentification**
- **Cloud Exploits (Azure, AWS, GCP)**
- **Firmware / ICS / OT**
- **Crypto-laundering**
- **Social psychology / Influence**

1 Are we prepared for an attack that does not exploit any technical flaws?

Example of a deepfake or a chat interaction with a supposed trusted person. Who will be able to detect it, and when? Who will know what to do?

2 Who holds the truth within the company?

If a deepfake or fake report comes out, who is legitimate to say '*it's fake*'? **How quickly can s/he respond?** Who will believe them? What false arguments might be put forward?

3 With the current AI tools, who in (or out of the company) **could fabricate a credible fake?** A disgruntled former employee, a frustrated employee, a hacktivist intern? **What competencies would be needed?**

4 Are our internal validation processes AI-proof?

If an internal AI automatically publishes incorrect/sensitive content, do we have a safeguard system in place? If an employee mistakenly publishes sensitive data, do we have the right to veto it?

5 How long does it to counter the fake?

If a false report is published, how many hours does it take to prove that it is fake without losing trust? In the world of AI, verification time becomes the weak link. Attackers exploit this inertia.

6 Who in the company knows how to spot image, sound or document manipulation? Do we have a good overview of possible manipulations? Beyond technical training and awareness, who knows how to manage an orchestrated credibility attack?

7 Do we have a strategy of credible allies to counter disinformation? Which trusted third parties (journalists, partners, recognized experts) **can relay a denial quickly and effectively?** The war of influence is won before the crisis by building a network of external trust.

8 Are our visuals, videos & official voices time-stamped, signed and traceable? Authenticating official content must be stored in an authenticity vault to prove the authenticity of official content in case of deepfake.

Data Poisoning and Model Manipulation

Oz'n'go

The attacker seeks to **undermine the integrity or credibility of the AI** the company uses for critical decisions.

He **targets the training data or the model's internal logic**, corrupting the AI's 'memory' from within.

It is an attack on the training or internal logic of AI that affects the company's strategy.

1 **Can we rapidly prove a model has been compromised, trace the source** (internal, external, supplier), **and assess the impact on our legal liability and insurance?** Upcoming legislation (AI Act, DORA, etc.) requires training data to be traced and audited.

2 **Do we monitor the data feeding our models and the automated retraining processes? Do we have a mechanism that detects dataset contamination or model anomalies in time?** Most companies monitor their servers, not their models.

3 **Do our teams understand that AI systems can be manipulated without being detected, and that their behaviour may shift with no technical warning? Are they able to recognise the weak signals of a compromised model?** Vigilance must become part of our culture: always verifying the AI's output before acting on it.

4 **If our pricing, production, or communication AI is compromised, how long would it take for us to detect it? Do we have a non-AI fallback plan? Do we know how to freeze a contaminated model during a crisis?** Many companies are dependent on AI systems without knowing how to regain control.

5 **Can we explain to a third party an incorrect decision produced by a contaminated model? Do we have the ability to audit and trace the training data to demonstrate manipulation in the event of an investigation or public crisis?** The risk is reputational, ethical, and trust-eroding.

6 **Who controls our external data sources?**

Modern poisoning often originates from a supplier, an open dataset, a pretrained model, or an external AI agent. No organisation fully controls its entire chain.

Model Inversion and Extraction

The attacker **targets what the model knows,** what it learned during training and simulations.

He also **attacks the model's memory** to reconstruct sensitive information. He also attempts to **steal invisible intellectual property** by attacking the model itself (its logic, structure, parameters, behaviour).

Objectives: resale of data, sabotage of trust in the model, ransom, industrial espionage.

This is an attack on business operations – directly impacting confidentiality and the company's core assets.

1 **Which of our AI models, agents, or internal tools represent genuine intellectual property, and which ones can be queried externally** (APIs, assistants, bots, agents), **even indirectly?** Identifying and classifying AI assets is essential to determine which are most exposed or critical, and to ensure adequate technical and legal protection.

2 **What would the loss of a model mean for your business and reputation in the short, medium, and long term?** Impact analysis helps clarify risks and prioritise the measures needed to implement business continuity and appropriate reputational safeguards.

3 **How would we manage a crisis where a stolen or cloned model is used to generate content that appears to come from us?** A governance of truth and a clear verification strategy become crucial to maintain the confidence of customers, partners, and the public.

4 **Do we have the technical evidence required to prove that a stolen or modified model is not ours** (signatures, hashing, secured versioning, logs, cryptographic fingerprints)**?** This is essential to issue credible public statements and defend the company's position.

5 **If a stolen model leaks sensitive data or if a model inversion reveals training data, can we identify the source of the leak, assume the right level of responsibility, and comply with regulatory obligations** (AI Act, GDPR)**?**

6 **Are our exposed AI endpoints protected and monitored to block model extraction attempts?** Do we control the volume, patterns, and nature of incoming queries (APIs, agents, bots)? Do we apply throttling*, segmentation, and behavioural monitoring to detect extraction attempts?

*throttling: a mechanism that limits or slows down the volume of requests sent to a system to prevent abuse, overload, or malicious data extraction.

Adversarial Machine Learning (AML)

The attacker becomes **an AI manipulator** — altering your models to distort the decisions they influence.

He tampers with sensors, warp perceptions, and induce incorrect choices.

Their **target is the AI model itself**, to deceive the model's perception.

This is an attack on strategic business operations.

1 **Which critical decisions are made by which AI model? Which models make automated decisions without human validation?** Have we mapped the models operating autonomously? If the AI decides alone = high risk. If humans re-validate = only disruption.

2 **Through which channel can a poisoned input** (image, sound, text, file, sensor, API) **deceive our AI, and can we replay the decision to understand how it was manipulated?** An AML attack often uses a barely detectable element. We need to trace the incident to reconstruct the full decision sequence and identify what must be corrected.

3 **Is our model trained to recognise when it is being manipulated? Do we have a protocol for adversarial testing?** Few organisations test their model against adversarial perturbations, verify robustness, or conduct red-teaming on their AI systems.

4 **Do we have a "buffer zone" to confirm a suspicious decision? Who detects abnormalities? Is there an alert threshold, and who is authorised to deactivate a compromised model?** An AML attack produces surprising, incoherent or statistically improbable outputs. Do we have a "stop-the-line AI" protocol?

5 **Do our teams recognise an aberrant output… or will they execute and apply the model's answer? Does our internal culture include anti-AI vigilance?** This is a frequent blind spot, where teams tend to trust the model too much.

6 **Do we know how to communicate, internally and publicly, about an AML error without triggering panic? Do we have a narrative ready for an "AI error"?** This is a crisis in itself, whether caused by misdiagnosis, corrupted pricing, or biased recommendations.

AI Supply Chain and Dependency Attacks

The attacker exploits the hub-and-spoke structure of a central provider and leverages slow detection and diluted responsibility to multiply victims before the alert is raised.

The fix is industrialised and deployed after the compromise.

The target is the invisible layers — multi-tenant cloud infrastructures, poorly secured connected devices, data pipelines, unsupervised autonomous AIs.

This is the logic of *"one-to-many attack"* and *"single point of compromise"* with massive propagation.

1 Where is our data located? How many AIs or AI agents manipulate it? How do we validate the reliability of the data feeding our AIs? **Have we mapped decision-making models and data pipelines end-to-end?** We must identify the invisible surface, prioritise components and zones to prevent contamination.

2 **Do we have cross-system visibility (IoT, AI, cloud) to detect abnormal behaviour in time? Do our AI agents have guardrails, or can they operate without human supervision?** We must move from local monitoring to systemic observation to assess the risk of uncontrolled autonomy.

3 **Who controls security across our cloud chain (multi-tenant or not)? The provider, us, or a third party?** If one link in the chain fails, how do we maintain business continuity? We must clarify contractual responsibilities and anticipate the propagation of an outage or manipulation.

4 **Are our service providers held to a patching schedule as strict as ours? How do we verify that announced patches were applied across the entire dependency chain?** We must anticipate the propagation of an outage/ manipulation and check upstream & downstream dependencies to prepare communications early.

5 **Can we clearly explain whether an outage or manipulation originated from the AI supply chain? Is our crisis communication aligned with cloud/AI providers in case of an incident? Who speaks first if one of us stays silent or contradicts the other?** This communication must be prepared in advance.

6 **Do we have a procedure to inform clients, partners and regulators, even when the root cause lies outside our control? Do we have a narrative ready for: "You should have better controlled your AI supply chain"?** In an AI supply chain attack, the company remains accountable, even if the origin is external.

THE NEW RISK IS INVISIBLE

With internal pressure as an expanding attack surface

- Under pressure, teams become more manipulable, less vigilant.
- AI amplifies exactly the destabilising signals: confusion, overload, doubt.

Hackers exploit what your management weakens

SIX IMMEDIATE LEVERS

The essential foundations of a modern defense

Truth Governance & Evidence Strategy

Define how the organisation establishes what is true vs. doubtful, who arbitrates, and how.Protect and authenticate sensitive official content.

Internal Alert & Critical Intelligence Channel (SME-friendly)

A simple pathway to report incidents, weak signals, AI inconsistencies, or online findings. Aggregate, filter and circulate useful info, even without a dedicated cyber team.

Expanded Resilience: Human, Cultural & Decision Reflexes

A psychological and systemic approach to build vigilance, cohesion and reliable reflexesunder uncertainty, doubt or manipulation.

Mastering Communications Under Pressure

Train executives and spokespersons to handle doubt, confusion and disinformation. The first minutes set the tone for the entire crisis.

Real-World Tests & Simulations

AI scenarios, fake messages, deepfakes, vishing, contradictory signals, etc.You don’t discover your crisis plan on the day of the crisis.

Critical Thinking & Anti-Manipulation Reflexes

Train teams to detect cognitive red flags, react without panic, use protection reflexesand report anomalies factually.

A prepared leadership, without a massive cyber budget.

WHO IS OZ'N'GO

Your Partner in Crisis Communication & Cyber Resilience.

Maryse Rebillot

Founder
Resilience & Crisis Communication

- 20+ years in marketing-communication in High Tech, Cyber security
- Expert in Crisis Management & Sensitive Communication
- International Executive Marketing Master, INSEAD

linkedin.com/in/maryse-rebillot

Experts Network

(mobilisables selon les missions, en fonction des besoins)

Cybersecurity

Cyber defense, White Hacking, Cyber Governance, Data Governance, Legal & Insurance

AI & Technologies

Gen AI, LLM, Vertical AI, Other AI

Human & Communication

Psychology, Medias & Journalists, Sociology

Oz'n'Go, the guide for leaders to master the crises AI creates, amplifies, or manipulates.

IN THE AGE OF CYBER-AI,
THE ONLY SUSTAINABLE DEFENSE

Is Human Cohesion

Maryse Rebillot

contact@ozngo.com

+41 77 533 77 67

www.ozngo.com