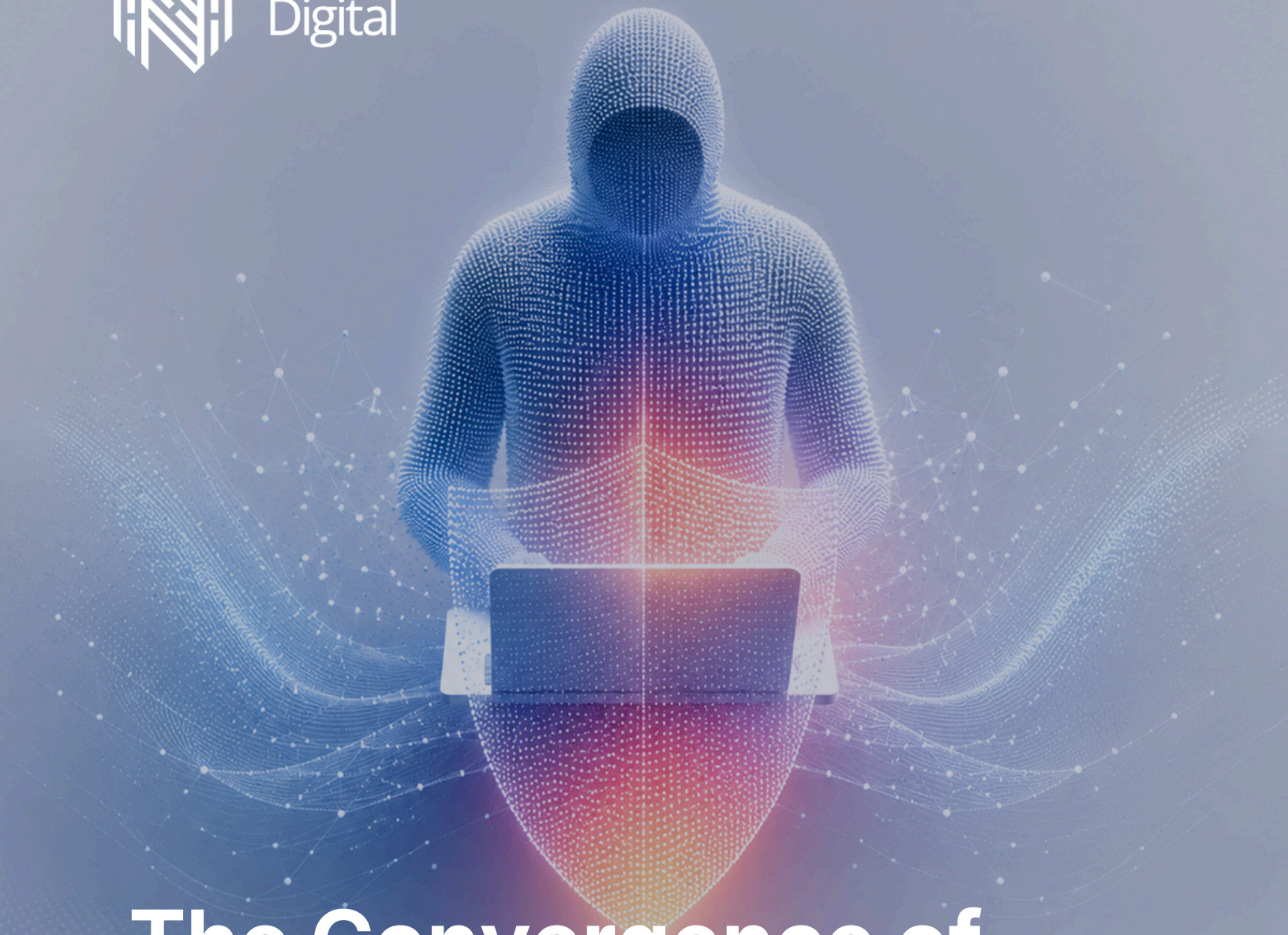
Nemko Digital



The Convergence of AI and Cybersecurity:

Why the Boundaries No Longer Exist

Two Disciplines, One Problem Space

For years, artificial intelligence and cybersecurity evolved as largely separate disciplines. Cybersecurity professionals focused on protecting systems, networks, and data, while AI practitioners concentrated on improving prediction accuracy, automation, and decision-making capabilities; That separation no longer exists.

Today, AI is simultaneously becoming a target of cyberattacks and a powerful capability for cyber defense. Organizations are deploying AI faster than they can secure or explain it, creating a new class of risks that traditional security frameworks were never designed to address. At the same time, attackers are leveraging AI to automate vulnerability discovery, accelerate exploit development, and operate at a scale previously impossible for human adversaries alone.

This intersection can be understood through two complementary perspectives:

- **Security for AI** – protecting AI systems against cyber threats.
- **AI for Security** – using AI to enhance cybersecurity capabilities.

The future of digital resilience will depend on mastering both.

Security for AI: Securing the New Attack Surface

Traditional software executes predefined logic. AI systems, however, learn from data and continuously make probabilistic decisions. This fundamentally changes the threat landscape.

Modern AI systems face threats that do not exist in conventional applications:

- **Model extraction and theft:** Attackers may replicate proprietary AI models through repeated querying, exposing valuable intellectual property and eroding competitive differentiation.
- **Prompt injection attacks:** Malicious instructions can manipulate model behavior, bypass guardrails, influence decision making, or trigger unauthorized actions across connected systems.
- **Training data poisoning:** Compromised training datasets can introduce hidden vulnerabilities, bias model outputs, and undermine the reliability of AI driven decisions.
- **Adversarial manipulation:** Carefully crafted inputs can deceive models into producing incorrect predictions, creating risks for critical business and operational processes.
- **Membership inference:** Threat actors may infer whether specific individuals or records were included in model training, creating privacy, confidentiality, and regulatory concerns.
- **Model inversion:** An attacker attempts to reconstruct sensitive information about a model's training data by analyzing its outputs or parameters.

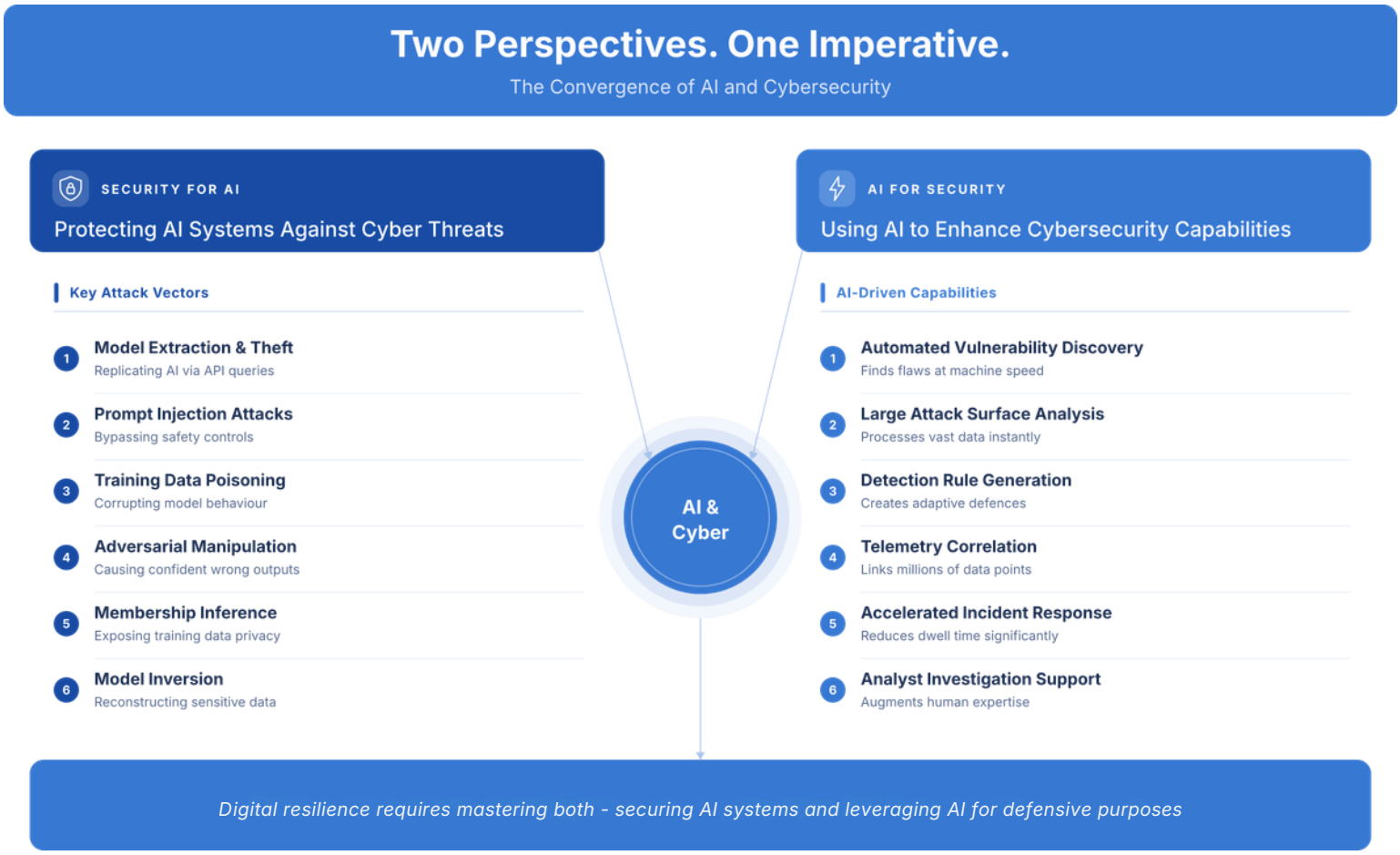


Figure 1: *6 Unique Threats AI Systems Face and 6 Capabilities that AI Systems bring to Cybersecurity.*

Organizations increasingly expose AI through APIs, cloud services, and shared platforms, creating opportunities for attackers to steal intellectual property or reconstruct sensitive information. Research has demonstrated that even protected learning-based models can often be replicated through carefully designed queries, creating both competitive and regulatory risks.

These risks are particularly relevant because AI is rapidly becoming embedded within critical infrastructure, industrial systems, healthcare environments, and connected devices.

Consequently, AI security is moving has moved from an academic concern to a business necessity.

Frameworks such as the OWASP Top 10 for Machine Learning, OWASP Top 10 for LLMs, and emerging Agentic AI guidance provide practical foundations for identifying and assessing these risks. In parallel, the EU AI Act increasingly emphasizes robustness and resilience requirements for high-risk AI systems.

The critical question is no longer whether AI can be attacked: it is whether organizations can realistically demonstrate that it has been tested sufficiently, with compliance requirements in mind.

AI for Security: Defending at Machine Speed

While AI introduces new risks, it also offers unprecedented defensive capabilities.

Cybersecurity has historically been constrained by human speed. Analysts investigate alerts manually, vulnerability researchers work sequentially, and threat hunting often depends on limited personnel resources. AI changes this equation, as advanced AI-driven security tools can:

- **Discover vulnerabilities automatically:** AI can continuously identify software weaknesses and misconfigurations at a scale and speed that would be impossible to achieve through manual security testing alone.
- **Analyze large attack surfaces:** AI can rapidly assess thousands of assets, applications, APIs, and configurations to uncover potential attack paths across complex environments.
- **Generate detection rules:** AI can assist security teams in creating and refining detection logic for emerging threats, helping organizations respond faster to new attack techniques.
- **Correlate massive volumes of telemetry:** AI can connect signals from logs, alerts, endpoints, networks, and cloud services to identify suspicious patterns that may otherwise go unnoticed.
- **Accelerate incident response:** AI can help prioritize, investigate, and contain security incidents more efficiently, reducing response times and limiting business impact.
- **Assist security analysts in investigations:** AI can augment human analysts by summarizing evidence, identifying relationships, and highlighting likely root causes, allowing teams to focus on higher-value decisions.

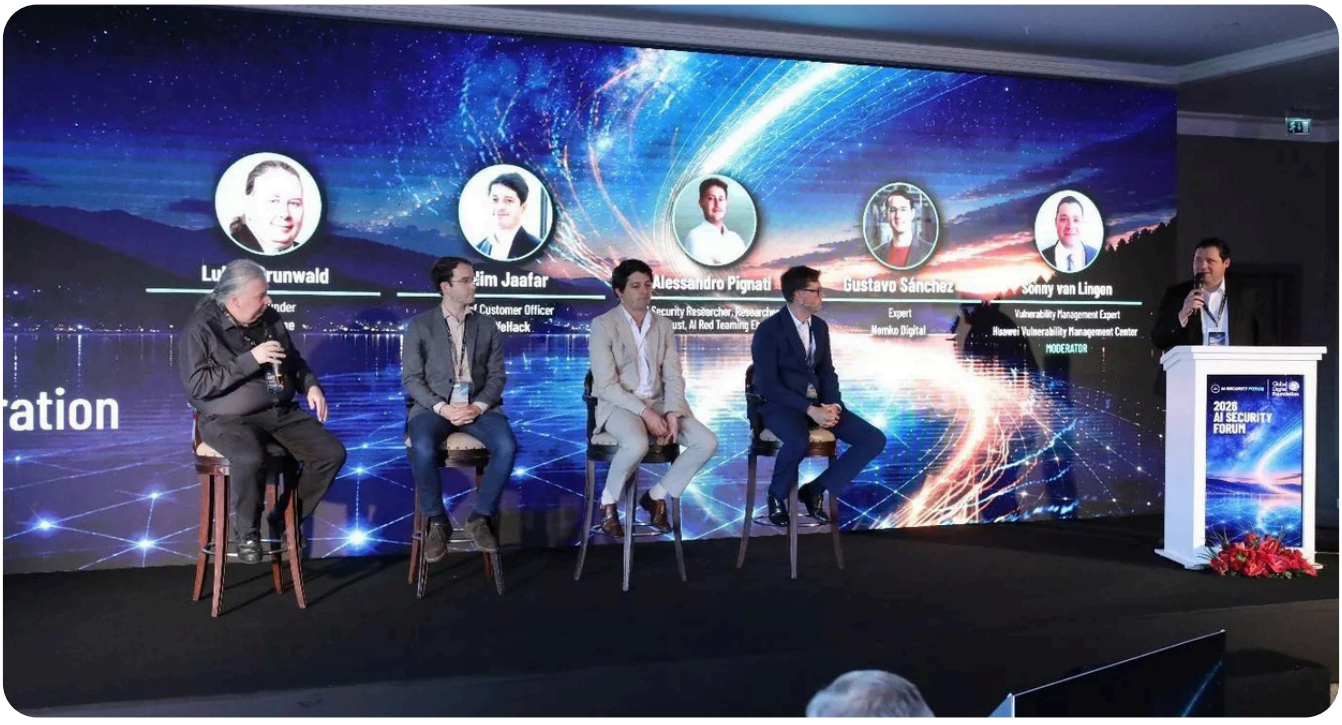
The result is a shift from human-speed security to machine-speed security.

Anthropic's Project Glasswing offers a glimpse into how profoundly AI may transform cybersecurity in the coming years. The initiative was launched to help secure critical software before increasingly capable AI systems can be leveraged by malicious actors. Within its first weeks, participating organizations used Anthropic's Claude Mythos model to identify **more than 10,000** high- or critical-severity vulnerabilities across widely deployed software and infrastructure. Several organizations reported finding vulnerabilities at a rate far beyond what traditional security teams could achieve, shifting the limiting factor from vulnerability discovery to verification, disclosure, and remediation.

Unfortunately, attackers benefit from this transformation. The same AI capability that helps defenders identify vulnerabilities can also help adversaries discover them. The same generative models that assist security teams can generate exploit code, automate reconnaissance, and adapt attacks dynamically. AI-driven threat detection tools are already being used to discover vulnerabilities and scale offensive operations faster than traditional approaches.

This creates an AI arms race where both sides continuously improve through automation.

Organizations that continue to rely exclusively on manual security processes risk becoming increasingly outpaced by adversaries. To stay ahead, establishing comprehensive AI management systems is crucial for maintaining an active defense posture.



Gustavo Sánchez (second from left) among other experts in a panel discussion at the 2026 AI Security Forum in Sapanca, Türkiye.

Explainability: Transparency versus Security

One of the most overlooked examples of the AI-cyber convergence is explainable AI (XAI).

Regulators increasingly demand transparency into AI-driven decisions. Organizations must understand, document, and justify how AI impacts users, especially under frameworks such as the EU AI Act. Explainability therefore becomes an essential compliance capability.

Yet transparency creates a paradox: The more information released about a model's behavior, the easier it may become for attackers to understand, manipulate, or replicate the system.

Explainability artifacts can provide valuable insight into how an AI system operates, but they may also expose information that attackers can exploit. By revealing key decision factors, model sensitivities, and structural weaknesses, explanations can help adversaries better understand how a model behaves and where it may be vulnerable. In some cases, these artifacts can even provide information useful for model theft, reverse engineering, or adversarial manipulation, creating a delicate balance between the transparency needed for trust and compliance and the confidentiality needed for security.

Research demonstrates that explanations can sometimes facilitate attacks by exposing insights into internal behavior. This forces organizations to confront a difficult question:

Can an AI system become more exploitable by making it more explainable?

The answer is increasingly yes.

Consequently, explainability artifacts should be treated as security-relevant assets rather than simple documentation outputs.

A compelling example of the broader challenges surrounding XAI and trustworthy AI can be found in the City of Amsterdam's attempt to develop a "fair" welfare fraud detection system. The project was designed with many of the safeguards commonly advocated in responsible AI frameworks, including transparency measures, bias testing, expert consultation, and fairness-oriented model development. Yet subsequent investigations questioned whether the system could truly avoid unfair outcomes in practice and highlighted the persistent tension between technical optimization and real-world societal impact. The case demonstrates that building trustworthy AI requires more than protecting systems from cyberattacks or ensuring technical robustness: AI systems must also be continuously scrutinized for fairness, transparency, accountability, and unintended consequences.

Learning from Failure: The AI Incident Database

One of the most valuable resources for understanding the real-world risks of artificial intelligence is the Artificial Intelligence Incident Database (AIID), a public repository that documents cases where AI systems have caused, contributed to, or been associated with harm or near-harm events. Inspired by incident reporting mechanisms used in industries such as aviation and cybersecurity, the database is built on a simple principle: organizations can only improve the safety and trustworthiness of AI if they systematically learn from failures. The incidents cataloged span a wide range of domains, including biased decision-making, safety failures, misinformation, privacy violations, autonomous system errors, and AI-enabled misuse. For organizations deploying AI, the AI Incident Database serves as a reminder that robust AI assurance requires more than model performance metrics: it demands continuous monitoring, incident analysis, risk management, and organizational learning throughout the AI lifecycle.

Organization	Total Reported Incident Involvement
OpenAI	169
Facebook	144
Google	99
Meta	72
Tesla	54
Microsoft	53
Amazon	36

Table 1: Ranking of organizations by approximate involvement across different roles in the AI Incident Database (developer, deployer, implicated system, harmed entity, etc.), as of 2nd July 2026.
Source: AI Incident Database

The Regulatory Driver

The convergence of AI and cybersecurity is no longer driven solely by technical innovation, as it is increasingly driven by regulation.

The European regulatory landscape now combines requirements from multiple domains, including:

- EU AI Act
- Cyber Resilience Act (CRA)
- GDPR
- Data Act

Historically, AI governance focused heavily on fairness, transparency, and accountability, while cybersecurity concentrated on confidentiality, integrity, and availability. Emerging regulations are forcing these disciplines together.

Therefore, under this new paradigm:

- Security testing becomes part of AI assurance.
- Threat modeling becomes part of compliance.
- Explainability becomes both a governance and security concern.
- Red teaming becomes a regulatory expectation rather than an optional best practice.

The result is the emergence of a unified concept: Trustworthy and Secure AI.

The Accountability Challenge

Perhaps the most significant unresolved issue lies in accountability. Modern AI deployments often involve foundation model providers, cloud vendors, fine-tuning specialists, software integrators, device manufacturers, third-party security vendors, etc.

And therefore, the following questions arise:

- When an AI-enabled product fails, who is responsible?
- When an AI model is compromised, who performs the security assessment?
- When outputs leak sensitive information, who is accountable?

Existing governance and security models struggle to answer these questions clearly, particularly in multi-vendor ecosystems.

As AI supply chains become more complex, organizations will need explicit allocation of responsibilities for security testing, explainability validation, incident response, and regulatory compliance.

A New Operating Model for Digital Trust

The convergence of AI and cybersecurity should not be viewed as the merger of two technical domains. It represents the emergence of a new operating model for digital trust, where security, governance, compliance, and AI assurance become inseparable. For business leaders, the implications extend far beyond the IT department. AI is increasingly embedded in products, customer interactions, business processes, and security operations, meaning that failures can affect intellectual property, operational resilience, regulatory compliance, customer trust, and ultimately competitive advantage.

Organizations therefore need to move beyond traditional cybersecurity programs and adopt AI-specific capabilities. This starts with AI-specific threat modeling, which helps identify how AI systems can be attacked, manipulated, or misused in ways that do not exist in conventional software. It should be complemented by AI red teaming and adversarial testing to evaluate how models behave under malicious or unexpected conditions. As explainability becomes a regulatory and business requirement, organizations must also ensure the secure handling of explainability artifacts, recognizing that transparency mechanisms can themselves expose sensitive information. At the same time, governance frameworks must be aligned with emerging regulations such as the EU AI Act and the Cyber Resilience Act, supported by clear accountability structures across vendors, developers, deployers, and operators. Finally, AI-enabled products should be subject to continuous resilience assessments, acknowledging that AI risks evolve throughout the lifecycle rather than ending at deployment.

For executives, the lessons are straightforward. First, understand where AI is already used within the organization and which systems are most business-critical. Second, ensure that AI systems are integrated into specificcybersecurity, risk management, and compliance processes. Third, establish clear ownership for AI-related risks, testing activities, and incident response. Fourth, begin exploring AI-specific assurance practices such as adversarial testing, AI red teaming, and continuous monitoring before regulations or customers require them.

These are no-regret actions. Organizations do not need to wait for perfect standards, mature regulations, or the next major AI incident before acting. Those that start now will be better positioned to adopt AI confidently, demonstrate trustworthiness to regulators and customers, and respond effectively as the threat landscape continues to evolve.

About the author

Gustavo Sánchez

Gustavo brings strong expertise at the intersection of AI security, trustworthy AI, and assurance for high-stakes systems. With a research-driven mindset and hands-on experience tackling real-world challenges — from adversarial ML to critical infrastructure contexts — Gustavo helps customers build AI systems that are not only innovative, but also secure, reliable, and demonstrably trustworthy.

Nemko Digital