CCN-CERT

BP/36

Security Recommendations to Counter the Offensive AI Models

BEST PRACTICE GUIDE

JUNE 2026

ccn-cert
centro criptológico nacional

Published by:

Date of issue: June 2026

Index

1. About CCN-CERT, National Governmental CERT

CCN-CERT is the Information Security Incident Response Capability of the National Cryptologic Centre (CCN), which is attached to the National Intelligence Centre (CNI). This service was set up in 2006 as the **Spanish National Governmental CERT.** This service was created in 2006 as the Spanish National Governmental CERT and its functions are set out in Law 11/2002 regulating the National Intelligence Centre, RD 421/2004 regulating the CCN and RD 311/2022, of 3 May, which regulates the National Security Framework.

Its mission is therefore to contribute to improving Spanish cybersecurity, acting as the national alert and response centre that cooperates and helps to respond quickly and efficiently to cyberattacks and to actively confront cyber threats, including the coordination at the public state level of the various Incident Response Capabilities or Cybersecurity Operations Centres in existence.

All of this is pursued with the ultimate aim of achieving a more secure and trustworthy cyberspace, safeguarding classified information (as set out in Article 4.F of Law 11/2002) and sensitive information, defending Spanish technological heritage, training expert personnel, applying security policies and procedures, and using and developing the most appropriate technologies for this purpose.

In accordance with this regulation and Law 40/2015 on the Legal Regime of the Public Sector, CCN-CERT is responsible for managing cyber incidents affecting any public body or company. In the case of critical public-sector operators, the management of cyber incidents will be carried out by CCN-CERT in coordination with the CNPIC.

2. Offensive AI: a paradigm shift that forces us to rethink cybersecurity

The incorporation of artificial intelligence into the offensive domain is redefining the foundations of cybersecurity. Its use has ceased to be an emerging threat and has already become an operational capability integrated into real campaigns by criminal and state actors. Cases documented by industry and by intelligence reports show multi-million-euro frauds supported by deepfakes, the automated generation of exploits, and highly personalised mass campaigns.

Its impact does not lie solely in the appearance of new threats, but in the **ability to accelerate, scale and automate already-known techniques, drastically reducing the time** between the identification of a vulnerability and its exploitation. Automation makes it possible to move at machine speed from spotting an opportunity to an effective attack, which reduces the defensive window and shifts much of the operational advantage to the attacker.

The implications for the **public sector** are especially relevant because of the criticality of essential services, the sensitive nature of the information managed, the interdependence between public bodies and providers, and the growing connection between IT and OT environments. In this context, offensive automation can affect the

continuity of services, increase the exposure of critical systems, and demand greater institutional coordination, information sharing, traceability and control over the AI systems deployed.

In the face of this paradigm shift, adaptation must rest on clear foundations:

- **Back to basics:** reinforce essential controls such as identity management, segmentation, continuous monitoring, access control and the protection of critical assets.
- **Transform IT and OT processes:** build in security by design, accelerate vulnerability management and patching, automate validations, and adapt controls to operational environments with low update capacity.
- **Design resilient and secure-by-default systems:** reduce the attack surface, control dependencies, limit unnecessary connectivity, and prepare systems to operate under adverse conditions.
- **Use AI as a governed defensive actor:** incorporate automated detection and response capabilities under strict criteria of control, human oversight and traceability.

The response must not be limited to adopting new tools; it requires reviewing the security model as a whole. This means integrating cybersecurity and AI-governance reference frameworks, aligning compliance, technical security and operational control, and establishing mechanisms for continuous evaluation of systems, data, models and agents. **Defensive AI** must be deployed with security-by-design criteria, permission control, protection of data flows, permanent monitoring, and periodic review of the risk model.

In short, offensive AI does not represent an incremental evolution but a paradigm shift that demands moving towards a more continuous, automated, resilient and governed form of cybersecurity. Only those organisations capable of combining control, automation and continuous adaptation will be able to reduce the asymmetry against attackers that already operate with increasingly autonomous, scalable and hard-to-anticipate capabilities.

3. Overview of the offensive AI model

Offensive artificial intelligence can be understood as the use of AI systems — including generative models, language models and autonomous agents— to support, automate or carry out phases of the attack cycle, such as target reconnaissance, phishing generation, malicious-code development, the identification and exploitation of vulnerabilities, or operational decision-making.

More than a completely new threat, it represents a **capability multiplier**: it increases the speed, scale, precision and degree of autonomy of already-known offensive techniques, putting these capabilities within reach of a larger number of actors who can carry out their actions without much knowledge. In fact, **its main change lies in the compression of attack timelines**. AI makes it possible to discover, chain and exploit vulnerabilities in far shorter periods, which limits the reaction capacity of organisations that still operate with analysis, prioritisation, patching or response cycles designed for weeks, not hours.

Several sources point in this direction. According to Gartner, a year ago AI agents reproduced real vulnerabilities at rates below 10–30%. Today, models are close to **90%** with escalating attempts, autonomously discover zero-day threats, and make it possible to discover and exploit vulnerabilities without advanced knowledge. Previously, attackers' lack of knowledge was an additional barrier that provided indirect protection. AI reduces that protection to a minimum, meaning that, with its help, anyone can be an advanced attacker, which **generalises these capabilities and significantly broadens the number of actors who can carry out attacks. [1][2][3]**

The **PROMPTSPY** case shows how malware can integrate generative AI into its execution flow to make dynamic decisions based on the environment. This malware uses models such as Gemini to interpret the state of the system in real time, generate step-by-step action instructions, and automatically adapt to different devices or configurations. **[4]**

3. Overview of the offensive AI model

For its part, research by the **Institute for AI Policy and Strategy [5]** documents real campaigns in which AI agents carried out between **80% and 90% of offensive operations,** leaving the human in a supervisory role.

Along the same lines, bodies such as NIST and CISA warn that AI significantly increases the **speed, scale and sophistication of attacks**, affecting all phases of the life cycle and requiring new approaches to risk management **[6]**. ENISA also addressed the challenges posed by AI in 2023 through the publication of the document Artificial Intelligence and Cybersecurity Research **[7]**.

Artificial intelligence has consolidated itself as a strategic priority for organisations, with accelerated growth in both investment and operational deployment. According to Deloitte **[8]**, around **60% of employees already have access to AI tools**, reflecting massive adoption over a short period. However, this increase in use does not translate directly into maturity: only a limited share of organisations have managed to move from experimentation to industrialisation, and barely 34% are genuinely transforming the way they operate with AI. In this context, agentic AI emerges as the next phase of evolution, with projected adoption rates that could reach 74% over the next two years, up from 23% today. Nonetheless, this rapid expansion takes place in a setting of significant risk, since **only about one in five organisations has mature governance and control models** over these autonomous systems. **[5]** This combination of high adoption, growing impact and low maturity in control places agentic AI as a key opportunity and, at the same time, as one of the main emerging risks in cybersecurity, especially in relation to identity management, data access and the autonomous execution of critical actions.

Offensive AI does not introduce new threats: it collapses the timelines of the ones we already know. What protected us yesterday no longer protects us tomorrow, because the attacker now operates at machine speed, in parallel and without rest. **The response must not be limited to acquiring new tools; it requires redesigning processes [4].**

The AI-enabled attacker discovers, chains and exploits vulnerabilities in very short times. Basic hygiene ceases to be a good practice and becomes the only line of containment and defence.

Change-management, patching and procurement processes were designed for cycles of weeks and with a tedious preparation-and-action methodology. **Offensive AI operates in hours. If processes do not change, the tools do not arrive in time. [3]**

The asymmetry only closes if the security team operates with AI. But those agents are, in turn, a new attack surface that must be governed like any other critical system. **[5]**

3. Overview of the offensive AI model

[1] *Gartner*, https://www.gartner.com/en/documents/6579402

[2] *Thehackernews*, https://thehackernews.com/2026/05/hackers-used-ai-to-develop-first-known.html

[3] *Securityweek*, https://www.securityweek.com/the-zero-knowledge-threat-actor-and-the-end-of-responsible-disclosure/

[4] *Google*, https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools/
https://blog.google/innovation-and-ai/infrastructure-and-cloud/google-cloud/gtig-report-ai-cyber-attacks-feb-2026/

[5] *Institute for AI Policy and Strategy*, https://www.iaps.ai/

[6] *NIST - CISA*, https://www.nist.gov/itl/ai-risk-management-framework
https://www.cisa.gov/ai

[7] *ENISA*, https://www.enisa.europa.eu/publications/artificial-intelligence-and-cybersecurity-research

[8] *Deloitte*, https://www.deloitte.com/es/es/services/consulting/research/estado-ia-en-las-empresas.html
https://www.deloitte.com/content/dam/assets-zone2/es/es/docs/services/consulting/2026/state-of-ai-2026.pdf
https://www.deloitte.com/es/es/services/consulting/research/estado-ia-generativa-empresas-espana.html
https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-ai-in-the-enterprise.html
https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/articles/agentic-ai-insights.html

3.1. Current context of AI-based defence

Insecure use of AI in organisations

AI is being adopted by business lines faster than security controls mature, creating new attack surfaces. This mismatch generates risks associated with the ungoverned use of tools, models and agents, especially when they are incorporated into internal processes without clear criteria for validation, traceability or access control.

For this reason, those responsible for cybersecurity must implement a **validation model specific to artificial-intelligence systems**, based on recognised standards that make it possible to identify the risks inherent to these environments. Use references such as OWASP for AI applications and language models, MITRE ATLAS to simulate attack techniques, and NIST and ENISA guidance for continuous evaluation and risk management.

This approach must be put into practice through AI-adapted penetration testing, attack-simulation exercises, and recurrent validation of models and agents, integrated into the life cycle. In this sense, LabIA, **[1]** the artificial-intelligence challenge platform of the National Cryptologic Centre, is a very suitable tool for strengthening capabilities in security applied to AI models.

AI literacy

Those responsible for cybersecurity must develop a baseline knowledge of the new risks introduced by artificial intelligence, drawing on regulatory frameworks and official guidance that help them understand its real impact. In this regard, it is key to use as a reference the European Artificial Intelligence Regulation (AI Act), which establishes obligations on risk management, governance and oversight of AI systems, as well as to rely on standards such as the NIST AI Risk Management Framework and ENISA's recommendations on AI cybersecurity, which provide practical criteria for identifying, assessing and mitigating emerging threats. Furthermore, all of the above must be understood within the broader framework of legal technology-risk-management obligations established principally by two instruments: the NIS 2 Directive and the DORA Regulation. To these must be added the GDPR, since users frequently share personal data with the models.

Gartner **[2]** proposes the figure of the AI Expert within security teams to consolidate this knowledge. In the private sector, the emergence of a specific executive figure to lead the transformation associated with artificial intelligence is taking hold, known as the Chief Artificial Intelligence Officer (CAIO), conceived as a role equivalent in relevance to the CISO but focused on AI strategy, governance and risk. This position arises in response to the growing complexity

of AI adoption and the need to centralise its control, as organisations are moving responsibility for these systems up to the executive-committee level, with functions ranging from strategic definition to the oversight of the secure and responsible use of the technology **[3]**. In the public sector, this figure is mandatory in the Federal Government of the United States of America.

Security tools for AI

New specific technologies are emerging which, although still immature, will be key to discovering and controlling the use of AI, including Shadow AI.

The essential capabilities can rely on concrete solutions:

- **Protection of model interaction and use** such as Microsoft Azure AI Security / Prompt Shields and Lakera Guard, which make it possible to filter inputs, mitigate manipulations (such as prompt injection) and control the release of sensitive information.
- **AI life-cycle security (models, data and development)** such as Protect AI and HiddenLayer, which facilitate the identification of vulnerabilities in models and processes, as well as protection against manipulation and supply-chain risks.
- **Operational detection and response** such as Microsoft Sentinel, Elastic Security and CounterCraft, which make it possible to monitor, detect anomalous behaviour and respond to incidents, including advanced techniques such as deception for critical environments.
- **Governance, control and regulatory compliance** such as IBM watsonx.governance and the ENS, NIST AI RMF and AI Act frameworks, which facilitate use control, traceability, auditing and regulatory compliance.
- **Security of autonomous systems and agents** such as Alias Robotics, specialising in the protection of autonomous systems and agents, relevant in advanced-automation scenarios.

It is important to start planning their integration into the technology stack and into the operations of the security operations centre.

Governance as the foundation

The governance of AI will become a pillar of cybersecurity.

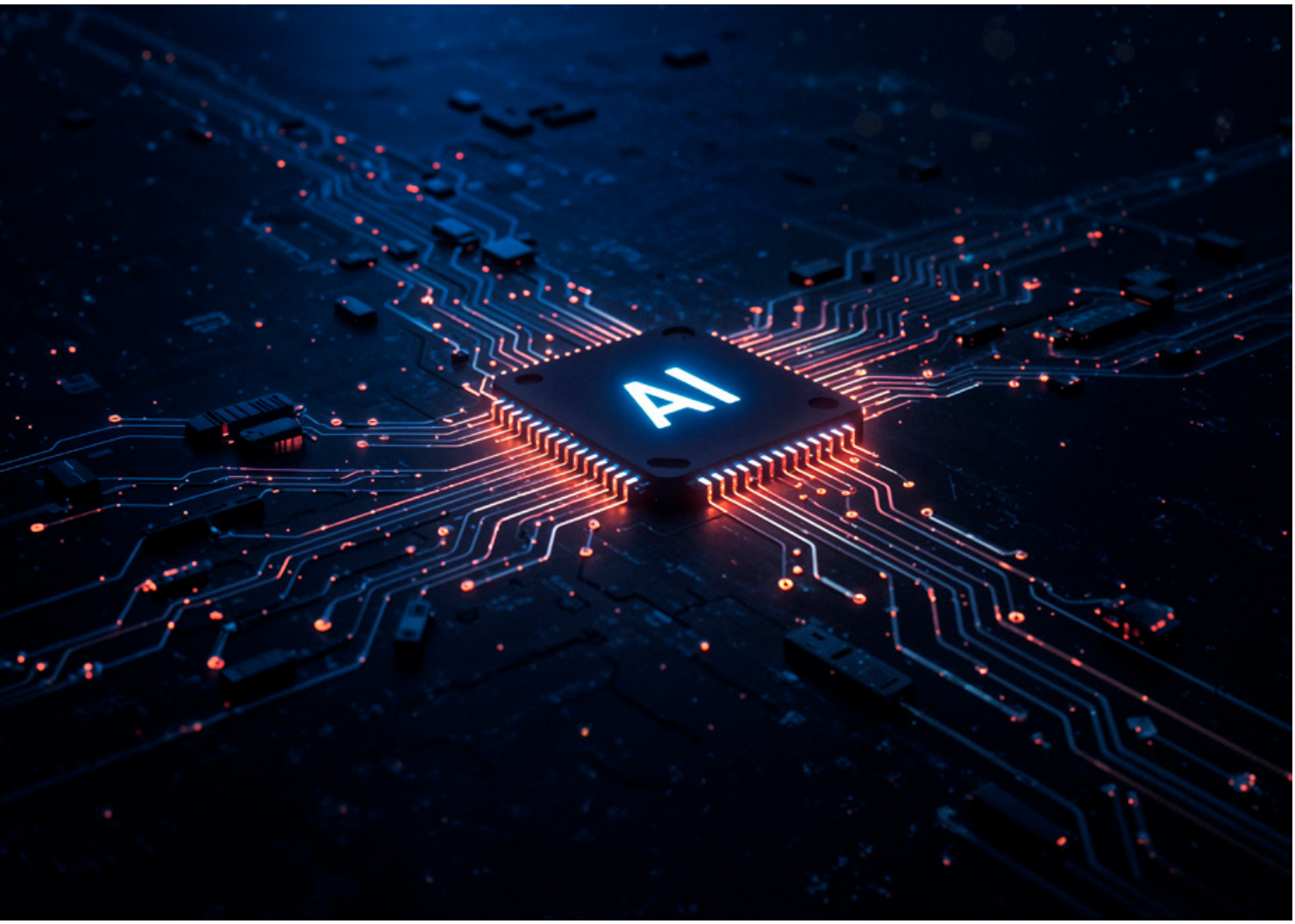
The management of the risk associated with the use of artificial intelligence must be aligned with the regulatory and governance frameworks defined at European and national level, rather than focusing solely on market approaches or vendor recommendations. In this sense, the implementation of controls must be based on the principles established by the European Artificial Intelligence Regulation (AI Act), which introduces specific obligations regarding risk management, transparency and human oversight, as well as on the cybersecurity guidelines defined by bodies such as ENISA, which stress the need to secure the entire

life cycle of AI systems. At national level, these recommendations must be integrated with the National Security Framework (ENS) and the CCN-CERT STIC guides, ensuring an approach consistent with the global management of information security. This combined framework makes it possible to structure protection against advanced threats, including those arising from offensive AI, through models of continuous control, dynamic risk assessment and effective oversight of the systems deployed.

[1] *LabIA*, https://labia.ccn.cni.es

[2] *Gartner*, https://www.gartner.com/en/documents/6579402

[3] *IBM*, https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/augmented-workforce



3.2. Threats, risks and possible scenarios in the use of AI

The adoption of artificial intelligence in cybercrime has moved from an experimental scenario to an operational one confirmed in real incidents. Intelligence reports and documented cases show that AI:

Reduces the technical complexity of attacks.

- The report by the Center for Long-Term Cybersecurity (UC Berkeley) indicates that generative models are lowering the barriers to entry, allowing actors with less technical knowledge to generate phishing, malware and reconnaissance tasks with ease. **[1]**
- The ENISA Threat Landscape 2025 report confirms that the availability of tools such as FraudGPT or WormGPT and the "phishing-as-a-service" model allow poorly qualified actors to run advanced campaigns, expanding the criminal ecosystem. **[2]**
- Academic research shows that AI-powered attacks reduce operational complexity by up to 72%, automating tasks that previously required high specialisation. **[3]**

Increases the success rate.

- The Microsoft Security Blog (2026) reports that the use of AI in phishing campaigns raises success rates to as much as 54%, compared with approximately 12% in traditional campaigns (≈450% improvement). **[4]**
- Academic studies also confirm that AI-assisted attacks achieve up to 67% greater effectiveness, thanks to their capacity for adaptation and personalisation. **[5]**
- ENISA notes that AI-assisted phishing is already dominant and more precise and effective, since it makes it possible to tailor messages to specific profiles and real contexts. **[6]**

Enables the automation of large-scale campaigns.

- The Google Threat Intelligence Group (GTIG) documents that state actors use AI to automate reconnaissance tasks, phishing generation and malware development at scale, integrating it across the entire attack cycle. **[7]**

- The ENISA 2025 report indicates that phishing is being carried out at "machine scale," with more than 80% of campaigns incorporating AI and being generated automatically. **[8]**
- Intelligence reports and global analysis show that attackers use AI to automate mass scans (up to 36,000 attempts per second) and to conduct reconnaissance and attack activity continuously. **[9]**, **[10]**

Identified threats

- **AI-driven advanced phishing:** an evolution towards highly targeted mass campaigns, even more adapted to the target, generated automatically and able to scale with high effectiveness, significantly increasing volume and overcoming traditional detection mechanisms.
- **Advanced impersonation via deepfakes:** use of AI to replicate the voice, image and behaviour of real people, enabling real-time attacks with high economic impact.
- **Generation of malware and malicious code (exploits) via AI:** AI is used to develop malicious code and discover vulnerabilities. For example, Google's cybersecurity team identified the first AI-assisted zero-day exploit, capable of evading second-factor authentication. This exploit was designed for a mass-exploitation operation. Characteristics typical of code generated by LLM models were detected.
- **Automation of open-source reconnaissance:** use of AI to collect and correlate public and technical information. AI is used to carry out faster analysis of stolen data, automatic profiling of victims and mass identification of targets.

In research by the Google Threat Intelligence Group **[11]**, it has been observed that attackers use AI as a "large-scale research assistant" to gather information about targets, identify the technologies used (technology stack, exposed services, versions) and rapidly analyse the potential attack surface.

One example is the use of AI by advanced actors to automate target reconnaissance before an attack, combining public sources with automated analysis. State actors and organised groups use AI models to analyse large volumes of public information (social networks, corporate websites, repositories) and generate detailed victim profiles and identify vulnerable organisations on a massive scale.

3. Overview of the offensive AI model

By combining open-source intelligence (OSINT) with direct reconnaissance capabilities over infrastructures, attackers can accurately map digital assets such as websites, programming interfaces, cloud services or infrastructure-as-a-service environments. AI also makes it possible to automatically correlate this data, identify the technologies used, detect vulnerable configurations and prioritise the targets with the highest probability of success. Reconnaissance ceases to be a limited preliminary phase and becomes a dynamic, permanent process that significantly increases the speed, precision and reach of attacks, reducing the effort required and expanding the real exposure of organisations.

- **Automated and adaptive social engineering:** AI-based automation makes it possible to launch millions of highly credible attacks in a short time, turning this vector into the main initial-access mechanism. Automated and adaptive social engineering is enhanced by AI, since it makes it possible to create malicious automated conversation channels (chatbots), generate personalised dynamic responses, and produce convincing human simulations, which helps attacks adjust to the context and to the interaction with the victim in real time.

- **Attacks on AI systems:** these are attacks aimed at manipulating systems based on language models. Their characteristics include hijacking the model's behaviour, exfiltrating information and carrying out unauthorised actions:
    - Exfiltrate context information.
    - Evade system restrictions.
    - Manipulate automatic decisions.
    - Alter document-analysis or RAG processes (Retrieval-Augmented Generation – an AI technique that connects a language model with private or external data sources).
    - Influence autonomous agents that have access to tools.

Documented cases such as Microsoft 365 Copilot **[12]** or Slack AI **[13]** demonstrate that an attacker can introduce hidden instructions into seemingly legitimate emails, documents or messages, causing the system to ignore its original rules and carry out unforeseen actions, such as accessing internal information or leaking it to external destinations without the user's knowledge.

This type of attack exploits a structural weakness of the models, which are unable to reliably distinguish between data and instructions, making it possible to hijack the system's behaviour, exfiltrate sensitive information and even trigger automated actions on connected systems. As a result, LLMs cease to be merely an information-exposure surface and become a critical operational-control point within the infrastructure, with direct implications for organisational security.

To these vectors is added a structural risk inherent to systems with persistent memory or retraining capabilities: the progressive accumulation of context between sessions.

Unlike memoryless models, these systems retain information from previous interactions that can be combined with new data to generate unforeseen connections between seemingly unrelated pieces of knowledge. An adversary who interacts deliberately and continuously with these systems can exploit that accumulated knowledge to identify patterns, infer sensitive information or chain actions that the defender —human or automated— had not anticipated. This risk is especially relevant in architectures with access to external knowledge bases —such as RAG architectures— or in environments where multiple agents share context. **[14]**

Agentic AI introduces a profound transformation in cybersecurity, from both an offensive and a defensive standpoint. On the one hand, it makes it possible to automate the complete attack cycle, from reconnaissance to exfiltration, through data analysis, contextual decision-making and autonomous execution, which increases speed, scale and adaptability while reducing the need for human intervention. In addition, the incorporation of these systems as "digital workers" opens up new risk vectors, as they can be manipulated or used to access critical information or carry out unauthorised actions.

On the other hand, these same capabilities enable an evolution towards more automated and proactive cybersecurity models, where agents make it possible to detect anomalies, analyse events, coordinate responses and manage risk in real time, improving operational

efficiency and reducing reaction times. However, this duality creates a key challenge: the gap between adoption and governance, since the deployment of these systems advances faster than the control mechanisms, introducing risks in terms of security, privacy and traceability.

In this context, the key lies in moving towards a model of governed autonomy, based on control, oversight and the limitation of capabilities, ensuring visibility over the behaviour of agents and maintaining human intervention in critical decisions. In short, agentic AI redefines the balance between attacker and defender, placing the advantage with those who are able to operate these systems with greater control, visibility and speed.

The recent evolution of artificial intelligence towards agentic models (systems capable of perceiving, reasoning, deciding and carrying out actions autonomously) is introducing a structural change in the field of cybersecurity. According to Deloitte's reports on the state of AI, organisations are accelerating the adoption of these systems, moving rapidly from pilot phases to production environments, which significantly broadens the risk surface if adequate control and oversight mechanisms are not in place. In this context, the appearance of advanced agents such as OpenClaw or Hermes evidences a relevant shift in orchestration capability, especially when combined with unrestricted local models, from which safety mechanisms have been removed and which respond entirely according to their training. This evolution facilitates the automation of traditionally complex tasks such as network scanning, vulnerability identification or even exploitation, making it possible to carry out the entire attack chain autonomously or semi-autonomously, with a significant reduction in human effort and a substantial increase in the speed and scale of operations.

---

[1] https://cltc.berkeley.edu/2026/02/11/new-cltc-report-on-managing-risks-of-agentic-ai/

[2] https://digital4security.eu/enisa-threat-landscape-2025/

[3] https://pdfs.semanticscholar.org/

[4] https://www.microsoft.com/en-us/security/blog/

[5] https://pdfs.semanticscholar.org/

[6] https://socket.dev/blog/enisa-threat-landscape-2025

[7] https://blog.google/threat-analysis-group/google-tag-generative-ai/

[8] https://www.cybersixt.com/enisa-threat-landscape-2025

[9] https://www.fortinet.com/blog/threat-research

[10] https://www.allaboutai.com/cybersecurity/ai-attacks/

[11] *Google,* https://blog.google/threat-analysis-group/google-tag-generative-ai/ https://blog.google/threat-analysis-group/updates-on-threat-actors-use-of-ai/ https://blog.google/threat-analysis-group/new-report-on-ai-and-security/

[12] *Microsoft 365 Copilot,* https://arxiv.org/abs/2509.10540 | CVE-2025-32711

[13] *Slack AI,* https://www.theregister.com/

[14] *AEPD,* https://www.aepd.es/guias/orientaciones-ia-agentica.pdf

Relationship of risks, examples and controls

RISK FAMILY	DESCRIPTION	EXAMPLES	RECOMMENDED CONTROLS
Offensive automation	Use of AI to accelerate reconnaissance, exploitation and lateral movement.	Mass reconnaissance, assisted exploitation, payload generation.	CTEM, ASM, KEV/EPSS, behaviour-based detection.
Synthetic social engineering	Use of AI for impersonation, phishing, fake voice or video.	Targeted phishing, deepfake, voice cloning.	Out-of-band verification, training, anti-fraud controls.
Attacks against AI systems	Manipulation of models, agents or RAG flows.	Prompt injection, data leakage, tool abuse.	Input/output validation, isolation, auditing.
Abuse of autonomous agents	Agents that carry out actions without sufficient control.	API calls, changes, data extraction.	Least privilege, human oversight, kill switch.
AI supply-chain risk	Dependencies, models, connectors, plug-ins, MCP, third parties.	Ungoverned components, shadow AI.	AI inventory, SBOM/AI-BOM, third-party assessment.

Possible scenarios

- **Accelerated exploitation of internet-exposed vulnerabilities.**
  An attacker detects a vulnerability published in a web service and exploits it within a few hours, before the organisation can apply the patch.
- **Hyper-personalised phishing against executives or privileged staff.**
  AI-generated emails containing real information about the recipient are sent to induce them to approve fraudulent access or transfers.
- **Voice or video deepfake for fraudulent authorisation.**
  A simulated video call or voice message from an executive is used to authorise payments or critical changes.
- **Use of AI for mass reconnaissance of public bodies.**
  An automated system analyses websites, documents and exposed services to identify vulnerabilities and access points.

- **Compromised AI agent carrying out unauthorised actions.**
  An agent integrated into internal processes is manipulated to access sensitive data or perform operations outside its function.

- **Provider affected by a vulnerability discovered and exploited with AI.**
  A flaw in a technology provider is automatically identified and exploited, cascading to multiple dependent bodies.

- **Information leakage through prompt injection or connector abuse.**
  A malicious input causes an AI system to access internal information and expose it in its responses or send it to third parties.

3.3. Strategic implications for the public sector

The recent evolution of artificial intelligence applied to cybersecurity, especially with the emergence of advanced models such as Claude Mythos Preview or GPT-5.5-Cyber, introduces a structural change in the balance between attackers and defenders. This change directly affects the security, resilience and governance of the public sector, transforming cybersecurity into a problem of speed, scale and continuous adaptation capacity.

The use of artificial intelligence makes it possible to carry out attacks at machine speed, drastically reducing intrusion and exploitation times. Intrusions can progress within minutes and attack cycles have been compressed from days or weeks to times close to immediate execution, reducing the response capacity of organisations.

- **Collapse of the vulnerability cycle**

  Artificial intelligence has transformed the traditional vulnerability cycle. Currently, identification and exploitation can occur almost simultaneously, eliminating the time window in which organisations could react before being compromised. This phenomenon renders classic vulnerability-management models obsolete.

  The appearance of advanced AI models specialised in cybersecurity represents a turning point. These models can identify vulnerabilities at scale, generate complex exploits and chain attacks autonomously. Their ability to find thousands of critical vulnerabilities in widely used systems evidences a qualitative change in global risk.

Vulnerabilities are no longer limited solely to new CVEs or zero-day vulnerabilities; the range also extends to insecure configurations, undue exposures and risks derived from third parties. AI-based automation of reconnaissance makes it possible to systematically identify these weak points across the entire attack surface, including misconfigured services, excessive access and permissions, or compromised external dependencies. As a result, the focus of risk expands from the code to the operational environment as a whole, increasing the probability of exposure and the complexity of managing it. All of the above also has an impact on regulatory compliance, since Regulation (EU) 2024/2847 of the European Parliament and of the Council of 23 October 2024, on horizontal cybersecurity requirements for products with digital elements, known as the Cyber Resilience Act, creates a new legal regime for vulnerability management and notification that enters into force in September 2026.

- **Short-term implications:** in the short term, a significant increase is expected in the number of vulnerabilities detected and exploited, an increase in automated attack campaigns, and a reduction in available patching times. Public administrations will have to face environments where exposure materialises almost immediately after a vulnerability appears.
- **Medium-term implications:** in the medium term, the industrialisation of cyberattacks will allow actors with lower technical capacity to carry out advanced attacks. Automation, scalability and the reduction of the attacker's operational costs will increase the systemic risk to critical infrastructures and public services.

Exposure of the human factor

Artificial intelligence amplifies the effectiveness of social engineering through techniques such as advanced phishing, deepfakes and personalised content generation. This reduces staff's ability to identify attacks, making the human factor an even more critical vector of compromise.

Impact on service continuity

The automation and speed of attacks increase the risk of disruptions to essential public services. The simultaneous exploitation of vulnerabilities in interconnected systems can generate cascading failures with a direct impact on citizens.

Impact on governance and regulation

The dual nature of these models forces a rethinking of regulatory and control frameworks. The management of capabilities such as Mythos becomes a strategic matter affecting national security, economic stability and the protection of critical infrastructures.

3. Overview of the offensive AI model

Paradigm shift in public-sector cybersecurity

The public sector must evolve towards continuous, automated and intelligence-based security models. The integration of real-time detection capabilities, automated response and the use of defensive AI becomes an essential requirement for maintaining resilience against threats that operate at machine speed.

3.4. Foundations for adaptation

>> The response must not be limited to acquiring new tools; it requires redesigning current processes to adapt them to the pace that the automation and agility of AI agents as attackers can impose.

BACK TO BASICS

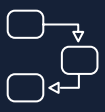
"What we already knew how to do, we now have to do at machine speed"

>> Cybersecurity is no longer just a knowledge problem, but a problem of execution speed, where controls must operate in real time so as not to fall systematically behind the attacker.

The AI-enabled attacker discovers, chains and exploits vulnerabilities in hours. Basic hygiene ceases to be a good practice and becomes the only line of containment.

BACK TO BASICS

"AI does not change our cyber strategy, but it does accelerate the need to strengthen the basics"



>> Current technology infrastructures are designed for human threats; they must be redesigned.

Change-management, patching, procurement and other processes were designed for cycles of weeks and with a tedious preparation-and-action methodology. Offensive AI operates in hours. If processes do not change, the tools do not arrive in time.

CHANGE IT PROCESSES

"Implementing defensive AI is not enough; we have to rethink the timing of the processes that surround it"

>> Accelerating processes seems sufficient, but it is not. Increasing scanning or patching speed does not solve the fact that validation, testing and deployment cycles continue to operate on incompatible timeframes, and forcing them entails risk and increases the likelihood of introducing failures greater than those being corrected. The problem is not only one of speed; it is one of design.

Organisations must accept that vulnerabilities will always exist and that the goal must be to make their exploitation difficult from the architecture, through defence in depth, environment isolation and centralised deployments. In this context, the competitive advantage will no longer lie solely in reacting faster, but in building systems that are resilient by default, where security does not depend solely on response time but on the very design of the environment.

RESILIENT AND SECURE-BY-DEFAULT DESIGN

"The future of cybersecurity does not depend only on going faster, but on redesigning systems to be resistant by default"

>> If the attacker uses agents, the defender has to use them too, but in a controlled way.

The asymmetry only closes if the security team operates with AI. But those agents are, in turn, a new attack surface that must be governed like any other critical system.

AI AS A DEFENSIVE ACTOR

"From AI as a risk to AI as a working tool"

4. Best practices

4.1. Best practices for IT processes

Traditionally static processes that must adapt to the new paradigm: vulnerability management, change management, the software development life cycle and application security, identity and access, backups and disaster recovery, and third-party risk.

VULNERABILITY MANAGEMENT

NOW:

In the current model, patching is planned in scheduled windows, prioritised by CVSS, and direct changes to production are avoided except in emergencies.

WHERE WE NEED TO GO:

We must move towards a model in which continuous scanning and patching is carried out, prioritised according to the real exploitability of vulnerabilities (KEV, EPSS). In this context, it is more important than ever to accurately identify critical assets, so that they are prioritised for review and patching, regardless of the existing layers of defence.

Likewise, it is key to enable auto-remediation mechanisms for bounded scenarios, carrying out corrective actions autonomously, but always within a well-defined set of use cases and with strict security controls (guardrails) that ensure control and prevent unwanted impacts.

CHANGE MANAGEMENT

NOW:

Each change goes through a process with forms, manual approvals and rigid deployment windows.

WHERE WE NEED TO GO:

In the future, changes travel as code in automated pipelines. Human decision-making only comes into play for genuinely critical changes.

This implies thorough documentation of changes, so that a rollback can be carried out with knowledge of the changes introduced.

SOFTWARE DEVELOPMENT LIFE CYCLE AND APPLICATION SECURITY

NOW:

The development model reviews security once per deployment, at the end of the cycle, and the development team experiences it as a formality.

WHERE WE NEED TO GO:

Each proposed change to the code must be analysed automatically by AI to look for weak points; an analyser reviews the code line by line and the system's component inventory is updated, in the background, without stopping work. Security shifts to the start of the cycle and ceases to be a blocker.

IDENTITY AND ACCESS

NOW:

Today the use of MFA is not widely extended; it sometimes uses SMS as the communication channel with the user, credentials are persistent without a rotation policy, and the review of permissions and access is a manual process.

WHERE WE NEED TO GO:

The evolution of the process must include phishing-resistant authentication (such as FIDO2), ephemeral access (such as JIT) and a living inventory that also includes the identities of AI agents.

BACKUPS AND DISASTER RECOVERY

NOW:

Currently, backup and disaster-recovery policies for critical services are tested, at best, once a year. Plans and strategies are static documents that do not evolve or get updated except after a serious incident that demonstrates their lack of operability.

WHERE WE NEED TO GO:

Recovery capacity takes on even greater importance. Backups must be immutable and isolated from the production network. Operational-resilience strategies must be tested with real exercises, on much shorter cycles.

THIRD-PARTY RISK

NOW:

Providers are assessed with periodic questionnaires, and the evidence they provide is not updated until the next review.

WHERE WE NEED TO GO:

Oversight must be continuous, relying on advanced telemetry and inventory artefacts such as shared SBOMs, and evolving towards continuous third-party risk-management models that allow dynamic tracking of exposure. Likewise, it is necessary to incorporate specific practices for managing risk in the AI supply chain, including the use of SBOMs extended to AI systems (AI BOM) to improve the traceability of models, dependencies and data. In parallel, contracts with providers must incorporate specific clauses on the use of artificial intelligence, including the assessment of foundation models, the management of risk associated with AI providers, and obligations of transparency, auditing and notification. All of this is aligned with the emerging requirements of the European AI Act, which reinforce the need for control over the AI value chain, data governance and the continuous oversight of the systems deployed.

4.2. Best practices for OT environments

Owing to the particularities of the world of Operational Technology (OT), it is advisable to take these environments into account when assessing the impact of offensive AI, as well as when defining mitigation measures specific to this domain.

- **Update the OT risk analysis** by explicitly incorporating offensive AI capabilities as a threat factor. Existing analyses under IEC 62443-2-1 or NERC CIP must be revised to reflect the lowering of the attacker's knowledge threshold and the acceleration of reconnaissance phases.
- **Define a perimeter attack-surface management policy:** continuous inventory of internet-exposed devices, acceptable-risk criteria, a periodic review process, and clear patching responsibilities for IT–OT border devices.
- **Evolve towards a model of continuous, contextualised discovery of OT assets**, capable of automatically identifying devices, communications, operational dependencies and the criticality of associated processes, extending visibility across the entire operational chain.
- **Integrate the offensive-AI threat into training and awareness programmes differentiated for OT personnel,** distinguishing their specific risk profile from that of a corporate IT environment.
- **Strengthen accreditation laboratories for OT environments,** considering specific analyses for AI-based threats.
- **Review security controls in the OT software and firmware supply chain,** including integrity-verification mechanisms and update procedures that reduce the exposure window of OT devices
- **Reinforce segmentation between IT and OT,** with particular attention to network-electronics equipment inside control networks.
- **Deploy technical passive-monitoring capability in OT networks:** probes and capture points that provide visibility over industrial traffic and inspection of the protocols in use in each environment, without interfering with process availability. This capability is the technical substrate on which detection capabilities are built.

Finally, a particularly relevant case in OT environments is that of devices which, due to their technical characteristics or operational restrictions, do not allow updating or patching. In these cases, the organisation must resort to compensating controls that reduce exposure without intervening on the device.

[Source] S2 Grupo

4.3. Best practices by cybersecurity capability

The current context of AI-driven threats requires evolving from a traditional cybersecurity approach based on static controls towards a model structured around operational capabilities, aligned with the dynamic and automated nature of the attack. In this sense, effective protection no longer depends solely on implementing isolated measures, but on the organisation's ability to detect, prevent, respond and adapt continuously in the face of a constantly evolving risk environment.

This section organises best practices according to these key capabilities, providing an integral and actionable view of security, from the proactive identification of vulnerabilities to automated incident response. This approach not only reinforces the security posture but also facilitates the prioritisation of initiatives, the allocation of resources and integration with regulatory frameworks such as the National Security Framework (ENS).

In this way, capabilities become the backbone of cybersecurity, allowing organisations to evolve towards a more resilient, agile model adapted to the speed and sophistication of current threats.

Self-assessment

Self-assessment of the level of protection against offensive-AI threats is a critical element for organisations, especially in the public sector, to understand their real exposure in a context where the speed and sophistication of attacks exceed traditional

defence models. This exercise makes it possible to identify gaps in capabilities such as detection, response or AI governance, and to prioritise measures before such weaknesses are exploited. To carry it out effectively, organisations must rely on structured processes such as AI-specific risk assessments, continuous reviews of the exposure surface, attack simulation on AI systems, as well as monitoring, behaviour-analysis and threat-intelligence tools that integrate real-time signals. In addition, they can develop internal evaluation frameworks based on standards —such as ENS, ISO or OWASP— adapted to AI, that make it possible to periodically measure their level of maturity and adjust their controls to the constant evolution of threats.

Use tools provided by CCN-CERT, such as the self-assessment form available in INES, to learn both the current state and the recommended actions to be carried out to improve the level of protection

Culture and training

Reinforce organisational knowledge about the impact of artificial intelligence on cybersecurity, ensuring that those responsible understand both its risks and its operational implications. Train cybersecurity teams in the new attack vectors, including the manipulation of instructions in models (prompt injection), impersonation through manipulated content (deepfakes), autonomous agents and automated exploitation. Train both technical and management profiles in the secure use of AI-based systems, incorporating their limitations and associated risks. Integrate this AI literacy within corporate awareness programmes and extend it to all users. Finally, develop internal capabilities that make it possible to evaluate AI models, tools and agents from a security perspective, ensuring controlled, informed use aligned with the organisation's risk framework.

- Awareness of AI-assisted cyberattacks, hyper-realistic impersonation and voice cloning.
- AI training: development of AI cybersecurity training materials.
- Create an internal AI Red Team and a permanent Vulnerability Operation as future capabilities.

Define AI security strategy and governance

Implement an AI governance model that makes it possible to control its use, risks and capabilities within the organisation in a structured way. Define clear AI-use policies, establishing which systems can be used, in which contexts and under which controls.

Classify systems according to their level of criticality and risk, applying proportionate measures in each case. Assign specific responsibilities throughout the entire life cycle of the systems (from development to operation and oversight) and ensure rigorous control of access to models, agents and data, preventing improper use or unauthorised exposure. Integrate this governance within the global cybersecurity framework, aligning it with existing controls and processes, and establish continuous-oversight and audit mechanisms that make it possible to detect deviations and ensure operation that complies with the defined requirements.

- Acceptable AI risk level and reference framework updated with a second line.
- Create an AI Committee or Secure Innovation Forum. Where this is not a viable or operational solution within the organisation, integrate AI management into existing bodies, such as Information Security Committees, Risk Committees or Digital Transformation Committees.
- CSA (Cloud Security Alliance) AI Controls Matrix integrated with providers and internally.

Protect and reduce patching times

Assume that the window between the publication of a patch and its exploitation is minimal, constituting an immediate operational risk that must be addressed as a priority. Apply urgent patching of any vulnerability included in the Known Exploited Vulnerabilities (KEV) catalogue, treating those with active, network-exposed exploitation as a security incident. Complement this management with a real-risk-based prioritisation approach, using models such as the Exploit Prediction Scoring System (EPSS) to order the remaining vulnerabilities. Reinforce this approach with the implementation of phishing-resistant multi-factor authentication based on FIDO2, eliminating weak methods such as SMS. Integrate the principles of least privilege and Zero Trust transversally, limiting access and continuously verifying identities and devices, and adopt models such as the Cloud-Native Application Protection Platform (CNAPP) to improve visibility and protection in cloud environments. Establish demanding service levels, ensuring the application of patches in less than 24 hours on internet-exposed systems and within short timeframes elsewhere, and complement this approach with continuous security-posture-management capabilities, both in AI (AI-SPM) and in SaaS applications (SSPM), to maintain permanent control of exposure. In this context, patching ceases to be a negotiable option and becomes a critical activity managed according to the operational impact of its deployment.

- **Phishing-resistant MFA:** FIDO2 is an open passwordless authentication standard developed by the FIDO Alliance and the W3C. It allows access to websites, applications and operating systems using asymmetric cryptography, biometrics or physical keys. It is the best barrier against phishing and account theft. Eliminate static credentials.

- **KEV patching in 24h:** priority correction of vulnerabilities included in CISA's Known Exploited Vulnerabilities (KEV) catalogue. These vulnerabilities are prioritised for urgent patching because they are already being exploited by attackers. Code reviews to identify vulnerabilities and AI-assisted SAST, which combines traditional static code scanning with language models to identify and correct vulnerabilities. AI acts as an assistant, prioritising alerts, reducing false positives and generating code suggestions for remediation. Identify critical assets to prioritise them over the rest.
- Size the vulnerability-management team or service to scale up by an order of magnitude.
- Incorporate the principles of least privilege, Zero Trust and CNAPP (cloud-native application protection).
- Use frameworks and tools to manage the security posture (AI-SPM, SSPM).

Detect vulnerabilities before deploying

Detecting vulnerabilities before deployment has become an essential pillar of cybersecurity in an environment where threats evolve at great speed and attackers use automated AI-based capabilities. In this context, the time available between the identification of a weakness and its effective exploitation has been drastically reduced, going from days or weeks to hours or even minutes, which practically eliminates any margin for action once systems are in production. For this reason, it is essential to adopt as a premise that any vulnerability reaching production environments will be discovered and potentially exploited by an adversary, which forces the defence to be anticipated from the outset.

To respond to this scenario, organisations must evolve towards an approach in which security is an intrinsic part of the development cycle, integrated early and continuously. This means incorporating advanced AI-supported vulnerability-analysis tools, capable of examining not only the source code but also configurations, dependencies and system behaviour. These tools make it possible to identify weaknesses that go beyond classic errors (such as logic vulnerabilities, misconfigurations or complex exploitation chains) that often go unnoticed by traditional mechanisms.

In addition to analysis, it is key to reinforce reaction capacity through the automation of vulnerability correction. This translates into the automatic generation of recommendations or even directly applicable patches based on the findings detected, especially those coming from techniques such as static analysis, which makes it possible to review code without executing it. This process must include human review of possible false positives to avoid changes that could entail risks greater than those

being mitigated. In this way, the time between detection and resolution is significantly shortened, reducing the risk of exploitation.

These capabilities must be fully integrated into the development life cycle, forming part of the usual software build, test and deployment processes. This not only makes it possible to detect and correct errors before they reach production, but also favours the adoption of a preventive and continuous security model, in which the identification and mitigation of risks is carried out constantly and automatically.

Taken together, this approach makes it possible to significantly reduce the attack surface, limit exposure to threats and adapt security to the current pace of cyberattacks, characterised by their speed, automation and capacity for dynamic adaptation. In an environment dominated by artificial intelligence, anticipation ceases to be an advantage and becomes an indispensable operational necessity.

Review everything already deployed

The identification and correction of vulnerabilities in code already deployed in production is a critical activity in the current context, where systems in operation represent the main target of attackers. Unlike the development phases, production code accumulates dependencies, configurations and historical changes that increase the probability of exposure to previously undetected weaknesses. For this reason, it is essential to establish continuous processes for reviewing and analysing software in operation, assuming that any exposed component may be the object of exploitation.

In this process, it is a priority to focus efforts on those components that present a greater risk surface, such as internet-exposed systems, modules that handle untrusted inputs, or functionalities linked to authentication, authorisation and auditing, whose exploitation can lead to improper access or significant compromises. These elements must be the object of specific, recurrent reviews, as they concentrate much of the operational risk and impact on the organisation.

Likewise, it is especially relevant to carry out audits of legacy code or code without active maintenance, which usually lacks recent reviews and may have been developed under already-obsolete security standards. This type of asset constitutes, in many cases, one of the main entry vectors for attackers, by combining exposure with a lack of continuous vigilance. Identifying this code and prioritising it in review plans is key to reducing overall exposure.

For this approach to be effective, it is necessary to allocate specific, sustained engineering capacity for the correction of identified findings. Detecting vulnerabilities is not enough; the organisation must be prepared to address them quickly and effectively, integrating these tasks into the usual cycles of software maintenance and evolution.

This means allocating resources, setting clear priorities and ensuring that corrective actions are carried out within appropriate timeframes, avoiding the accumulation of technical debt.

Taken together, this model makes it possible to advance towards continuous security in production, in which the review, identification and correction of vulnerabilities form part of the normal operation of systems, progressively reducing the attack surface and improving resilience against increasingly automated and sophisticated threats.

Design assuming future compromises and security breaches

Adopt a design approach based on the assumption of compromise, accepting that no system is completely invulnerable and that, at some point, an attacker may gain access. This paradigm shift forces the prioritisation of impact containment and resilience over absolute prevention, designing architectures capable of limiting the spread of an incident and reducing its operational consequences.

To this end, it is necessary to implement a Zero Trust model effectively, in which every interaction between systems, services or users is treated as potentially untrusted. This implies explicitly authenticating and authorising each request, regardless of its origin, avoiding reliance on traditional perimeters and ensuring continuous verification of identity and context. This approach makes it possible to significantly reduce the risk of lateral movement in the event of compromise.

Complement this model with the adoption of robust device-bound authentication mechanisms, so that access does not depend solely on credentials but also on trust in the device used. This drastically reduces exposure to impersonation and credential-theft attacks.

Likewise, replace the use of persistent credentials or static secrets with a model based on ephemeral credentials, with limited validity in time and generated dynamically as needed. The elimination of static programming keys and other long-lived secrets is fundamental to preventing their reuse by an attacker in the event of leakage or compromise.

It is important not to limit the strategy to the acceleration of operational capabilities, such as scanning or patching, since reducing times does not solve structural risks and can compromise system stability if essential controls are omitted. It is essential to adopt a resilience-based approach, aimed at making the exploitation of vulnerabilities difficult even when they exist. Normalising basic design concepts such as the isolation of critical components, centralised management of corrections and the implementation of defence in depth —designing a system where the combination of layers drastically

reduces the probability and impact of an attack— will be more effective against new threats than adding controls indiscriminately.

In this sense, although advanced models make it possible to automate and scale the management of cybersecurity, they do not replace expert judgement or governance; it is key to evolve towards secure-by-default systems rather than relying exclusively on response speed.

Taken together, this approach makes it possible to evolve towards a security architecture in which systems are prepared to operate even under conditions of compromise, limiting exposure, controlling access and ensuring greater response capacity against advanced threats.

Manage the attack surface in real time

Ensure visibility and continuous control of all exposed assets, understanding that any unidentified or unmanaged element represents a potential entry point for the attacker. To this end, establish and maintain an updated, dynamic inventory of all accessible hosts, services and programming interfaces, including both traditional infrastructures and cloud environments and components deployed in a decentralised manner.

Complement this control with an active strategy of decommissioning obsolete or unmaintained systems, eliminating those assets that lack a clear owner or that cannot guarantee adequate levels of updating and support. This process must be deliberately demanding, since the permanence of this type of system significantly increases the attack surface and operational risk.

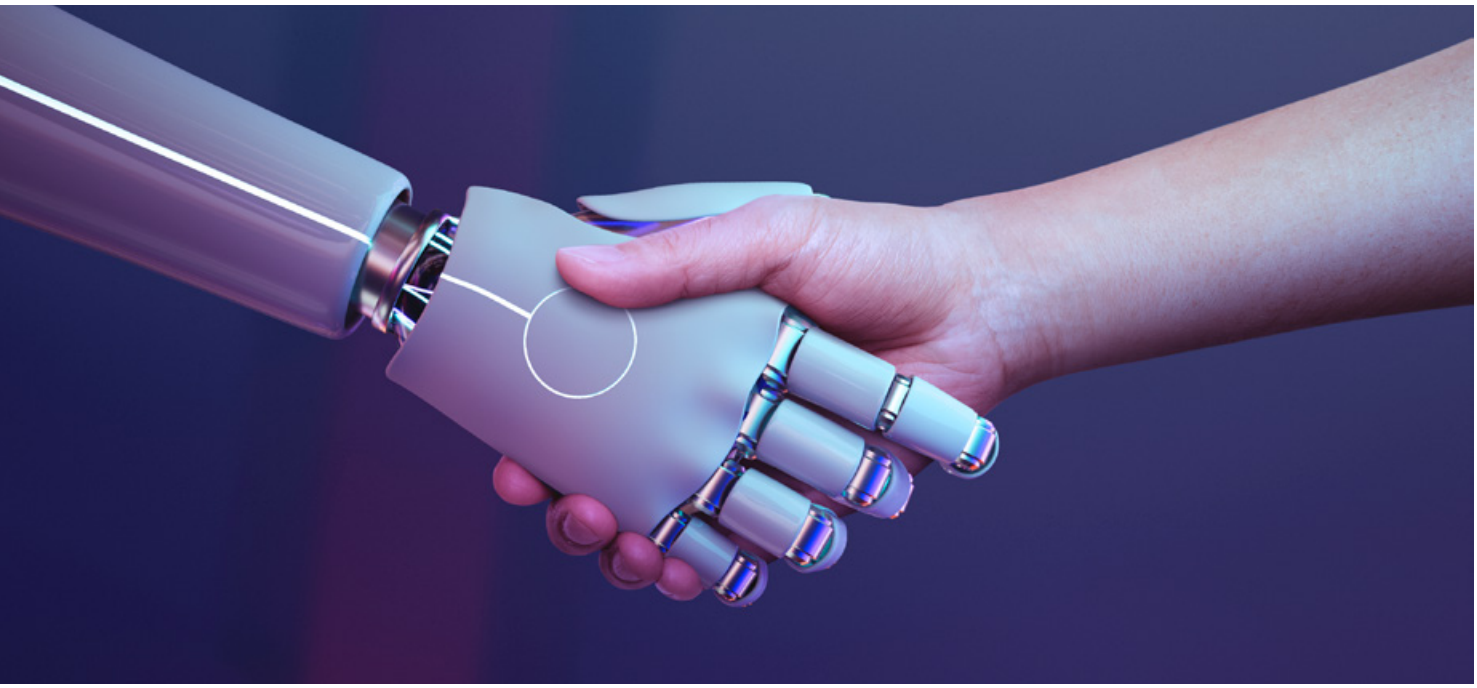
Additionally, apply attack-surface-reduction strategies, simplifying and implementing the principle of minimum necessary exposure, ensuring that each service or component offers only the functionalities strictly indispensable for its operation. This implies reviewing and limiting open ports, external access, unnecessary functionalities and superfluous dependencies, thereby reducing the attacker's ability to identify exploitable vectors.

Taken together, this approach makes it possible to evolve towards active management of the attack surface, in which the organisation maintains permanent control over its assets, reduces unnecessary complexity and effectively limits opportunities for compromise in an increasingly dynamic and distributed environment.

Improvement of detection and response capabilities

The reduction of incident-response times becomes a critical factor in the context of offensive-AI-driven threats, where attacks develop and escalate at machine speed, drastically reducing the reaction margin of organisations. In this scenario, delays of minutes or hours can mark the difference between effective containment and a complete compromise of the system, especially when adversaries automate exploitation and lateral movement. To ensure an effective response, organisations must evolve towards continuous detection-and-response models, in both IT and OT environments, relying on agentic-AI-based tools capable of correlating events in real time, identifying anomalous behaviour and carrying out automated responses, such as SOAR, XDR or autonomous SOCs. These capabilities make it possible to prioritise incidents, reduce manual intervention and activate immediate countermeasures such as system isolation, credential revocation or malicious-traffic blocking. Complementarily, it is essential to establish operational processes such as automated procedures, response simulations and periodic resilience exercises, integrating real-time threat intelligence to anticipate emerging attack vectors. Taken together, the combination of automation, contextual intelligence and orchestration makes it possible to significantly reduce response time and adapt it to the current speed of cyberattacks.

With regard to OT environments, two additional tasks beyond those already mentioned take on special relevance: characterising the normal behaviour of processes through learning from the operational history, and genuinely correlated IT+OT detection use cases that reflect the interdependencies already existing between the two environments.



4.4. Security in the use of AI agents

Artificial-intelligence agents represent a significant evolution over traditional AI systems, by incorporating capabilities for:

- Autonomous decision-making
- Execution of actions on systems
- Continuous interaction with digital environments

This new paradigm introduces relevant operational benefits, but also considerably broadens the risk surface, by allowing AI to act directly on critical systems without direct human intervention.

The main risk vectors arising from the use of AI agents are:

- **Risk vectors:**
  The main risks associated with the use of AI agents include the uncontrolled execution of actions, manipulation via prompt injection, information exfiltration, the chaining of erroneous decisions, the insecure use of external APIs, and a lack of traceability.

- **Security principles:**
  The secure use of AI agents is based on key principles such as least privilege, the limitation of autonomy, input validation, the separation of capabilities, human control over critical decisions, and continuous monitoring.

- **Technical controls:**
  Technical measures include access control, sandboxing, input/output filtering, secure credential management, control of calls to external APIs, and complete logging and traceability of the agent's actions.

- **Relationship between risks, principles and controls:**
  (See the table on the next page)

The use of AI agents requires a specific security approach focused on the control of actions, continuous oversight and governance. The adoption of these principles and controls is essential to ensure their secure use in critical environments.

RISK VECTOR	SECURITY PRINCIPLE	TECHNICAL CONTROLS
Uncontrolled execution of actions	Least privilege / limitation of autonomy.	Define strict access and isolated environments to limit the impact of AI agents: ● Implement Identity and Access Management (IAM) with minimal roles (Azure AD, AWS IAM). ● Use RBAC/ABAC to differentiate permissions by type of agent. ● Run agents in isolated environments: Docker containers with limited permissions; serverless cloud environments (AWS Lambda, Azure Functions).
Prompt injection	Input validation / separation of capabilities.	Protect the system against manipulation through malicious input: ● Implement data-validation pipelines before sending instructions (prompts) to the model. ● Use frameworks such as: Guardrails AI; Microsoft Prompt Shields / Azure AI Content Filters. ● Separate: system instructions; user context. ● Apply input sanitisation (allow-lists, detection of malicious patterns).
Data exfiltration	Information protection.	Prevent information leakage or unauthorised responses: ● Implement Data Loss Prevention (DLP) in responses. ● Content filtering with: Azure Content Safety; OpenAI moderation endpoints. ● Control data access through: intermediary APIs with validation; context minimisation (only necessary data). ● Use techniques such as output-validation rules (e.g. do not return credentials, tokens, sensitive data).
Chained errors	Human control / limitation of autonomy.	Introduce human oversight in critical actions: ● Mandatory approval flow for: configuration changes; access to sensitive data. ● Integration with tools enabling human oversight/validation. ● Define criticality levels: automatic actions (low risk); actions requiring approval (high impact).
Use of external APIs	Dependency control.	Control the agent's external interactions: ● Use an API Gateway. ● Apply policies: allow-list of endpoints; response limits; mandatory authentication. ● Monitoring to detect anomalous API use. ● Blocking of unapproved or out-of-pattern calls.
Credential management	Protection of secrets.	Protect secrets and avoid permanent exposure: ● Use secrets-management systems. ● Implement: ephemeral credentials (Just-In-Time access); automatic key rotation. ● Eliminate: API keys in code; secrets in prompts or code.
Lack of traceability	Monitoring and auditing.	Ensure traceability and incident-response capacity: ● Detailed logging of: prompts received; model decisions; actions executed. ● Integration with SIEMs. ● Event correlation to detect: anomalous behaviour; possible malicious-instruction injections. ● Retention and auditing for compliance (ENS, ISO).

4.5. Integration of reference frameworks in AI security

To ensure a sound approach aligned with international best practices, it is necessary to incorporate recognised reference frameworks that make it possible to structure security in AI-based environments coherently. These frameworks provide methodologies, controls and principles that facilitate both the implementation and the oversight of security measures adapted to the new technological context.

In this regard, it is advisable to adopt ENISA's "Multilayer Framework for Good Cybersecurity Practices for AI," which provides a clear structure based on three levels: general cybersecurity fundamentals, specific controls applied to AI systems, and measures adapted to each sector. This approach facilitates an integral view that combines traditional protection with the particular needs of artificial intelligence.

Likewise, the NIST AI Risk Management Framework constitutes a key reference for establishing models of governance, risk management and trust generation in AI systems. This framework is reinforced by more recent NIST publications, including specific profiles for generative AI and analyses oriented towards its use in critical infrastructures, which extends its applicability to more complex and sensitive environments.

In the area of technical risks, it is especially relevant to incorporate the OWASP guides for applications based on language models and generative-AI systems, which identify common vulnerabilities, such as input manipulation, information leakage or abuse of system capabilities. These references are fundamental for addressing emerging risks in AI-based applications and agentic systems.

In the Spanish public sector, the National Security Framework (ENS) constitutes the main regulatory and operational framework for articulating protection against AI-based or AI-assisted threats. International references —ENISA, NIST AI RMF, MITRE ATLAS, OWASP LLM/GenAI, ISO/IEC 42001 or Cloud Security Alliance— must be used as complementary instruments to enrich, update and specialise controls, but they do not replace the structuring function of the ENS or the role of the CCN as the technical reference body for its development, interpretation and practical application.

This integration also extends to private entities that use their own information systems and provide services or solutions to public entities within the corresponding scope of application, in accordance with the National Security Framework, approved by Royal Decree 311/2022 of 3 May.

Consequently, security against offensive AI must be integrated into the ENS risk analysis, the categorisation of systems, vulnerability management, information protection, traceability, continuity, provider management, incident detection and response, and continuous improvement. The emergence of AI-based offensive capabilities does not require creating a parallel framework, but rather updating the application of the ENS so that its principles, minimum requirements and security measures respond to a more automated, adaptive and rapid threat.

The integration of these frameworks makes it possible to build a robust security model that combines regulatory compliance with the adoption of international best practices in an environment marked by the growing influence of artificial intelligence.

4.6. Roadmap

FIRST STEPS

Close the patching gap. Phishing-resistant identity.

Critical systems prepared: no high or critical vulnerabilities and in continuous evaluation cycles.

Third-party review: begin to assess third parties from the perspective of protection against an AI scenario.

Second line: updated response plans and risk reference frameworks. Mitigation strategies.

>>

PROPOSED DESIGN

Pre-authorised emergency changes.

Vulnerability operations as a permanent function.

Detection at machine speed.

Agility in testing and acquiring defensive elements.

Resilient, isolated and secure-by-default design.

>>

BUILDING AGENTIC CYBERSECURITY

Own + external AI Red Team.

Governed defensive agents.

Defend your own agents.

Specific awareness of AI risks.

AI control framework.

>> FIRST STEPS - DETAIL

- **Close the patching gap:** prioritise patching of vulnerabilities on the CISA KEV (Known Exploited Vulnerabilities) catalogue within 24h. Use EPSS-based prioritisation for the rest of the vulnerabilities. Create pre-authorised extraordinary change windows.
- **Phishing-resistant identity:** the activation of multi-factor authentication must cover all remote access, with no exceptions. Where this is not technically feasible immediately, the situation must be documented as an accepted risk with a temporary compensating measure and a dated resolution plan. Migrate to authentication methods such as FIDO2/ passkeys, eliminate SMS-MFA, ephemeral tokens. FIDO2 allows you to authenticate securely using different methods:
    - **Personal device:** using your own computer or mobile via biometrics (fingerprint, facial recognition) or a local PIN. This is commonly known as passkeys.
    - **External physical keys:** USB, NFC or Bluetooth devices (such as YubiKeys or similar) that require physical presence and, often, the touch of a button or a PIN to approve sign-in.
    - **Main advantages:**
        - Phishing-proof (fake websites cannot fool the FIDO2 system; the key or device only responds on the official domain).
        - No master passwords.
        - Decentralised privacy (websites never store a password file that can be hacked).
        - Zero Trust in identity and device.
- **Critical systems prepared:** no high or critical vulnerabilities and in continuous evaluation cycles. The elimination of default credentials on network equipment must be verified, as well as ensuring the existence and accessibility of offline configuration backups for the most critical systems, with a documented minimum recovery procedure.
- **Third-party review:** begin to assess third parties from the perspective of protection against an AI scenario. This review must include an audit of providers' remote access, verifying that they operate under least-privilege principles and that sessions are audited.
- **Second line:** updated response plans and risk frameworks. Define mitigation strategies for vulnerabilities that cannot be patched within appropriate timeframes.

"AI does not change our cyber strategy, but it does accelerate the need to strengthen the basics."

>> REDESIGNS PROPOSED TO ADAPT TO AI THREATS

- **Pre-authorised emergency changes:** define who authorises, in which cases and within what times, something to be rehearsed in exercises and multi-scenario drills.
- **Vulnerability management as a permanent function:** continuous discovery, AI-assisted patching, continuous monitoring of the attack surface.
- **Detection at machine speed:** agents for the use cases, pre-authorised containment, incident-response procedures tested quarterly against an AI-enabled attacker. Response procedures must include specific procedures for the priority scenarios of the risk analysis, and response capacity must be sized to cope with the operating speed of offensive AI, with containment mechanisms that do not require manual approval in the most critical cases.
- **Agility in testing and acquiring defensive elements:** it is necessary to define agile processes for incorporating cyber capabilities. The speed factor introduced by offensive AI makes traditional evaluation and acquisition cycles obsolete; the validation and accreditation processes for new defensive capabilities must be redesigned to operate within timeframes compatible with the pace of evolution of the threat.
- **Implementation of defence in depth by design and architecture:** a security approach that consists of not relying on a single protective measure, but on establishing multiple layers of complementary controls that act in a coordinated way. The goal is that, if one of those layers fails or is overcome, others mitigate the impact, hinder the progression of the attack and protect critical assets. Vulnerabilities and failures are inevitable.

"Environments are designed where exploitation is complex, limited and detectable."

>> BUILDING AGENTIC CYBERSECURITY – DETAIL

- **Own + external AI Red Team:** prepare Threat-Led Penetration Testing (TLPT) exercises with AI. These exercises must incorporate simulations of social-engineering campaigns adapted to the specific context of each organisation, and their results must feed back into both the risk analysis and the training programme.
- **Governed defensive agents:** deployed agents must have a limited scope, a bounded blast radius, a kill switch, human oversight and intervention capability, and a defined, unique identity per agent. The goal is to have an AI-based defensive capability that effectively improves the prevention, detection and response to hostile operations carried out or supported by AI, with a progressive degree of governed autonomy.
- **Defend your own agents:** the adopted AI agents must be protected from threats such as malicious requests or orders (prompt injection), theft of the agent's credentials or tool manipulation, and audit logs must be created.
- **Specific awareness of AI risks:** the evolution of artificial intelligence has multiplied the effectiveness of attacks based on advanced social engineering, introducing increasingly hard-to-detect forms of impersonation. Hyper-realistic impersonation via deepfakes makes it possible to accurately recreate the image of executives or key employees in environments such as video calls, with a high level of credibility. Voice cloning makes it possible to reproduce complete conversations with realistic tone, rhythm and style, facilitating the fraudulent authorisation of sensitive operations. These capabilities are combined with AI-assisted phishing, which generates highly personalised messages tailored to the recipient's context, significantly increasing their success rate. In parallel, attackers can orchestrate campaigns specifically aimed at privileged profiles, such as senior management or financial officers, using contextual information and detailed profiles to maximise impact. Faced with this scenario, it is essential to reinforce verification processes through independent (out-of-band) channels, which make it possible to validate critical identities and requests outside the original channel, reducing the risk of impersonation and fraud. The training programme must be differentiated by profile, with practical exercises that include simulated phishing and vishing campaigns using terminology and context specific to each organisation. The frequency must be sufficient to keep pace with the evolution of the adversary's techniques.
- **Examples of AI control frameworks:** ISO/IEC 42001 establishes a management framework for AI systems, aimed at ensuring their governance, risk control and regulatory compliance. The NIST AI RMF complements this approach by providing practical guidance for identifying, assessing and mitigating risks associated with the AI life cycle, with an emphasis on reliability, security and ethics. ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) provides a taxonomy of threats specific to AI systems, useful for anticipating attack vectors. OWASP LLM/ASI identifies critical vulnerabilities and risks in language models and generative-AI systems, facilitating their protection; and, finally, MRM (Model Risk Management) extended to AI as a threat broadens traditional model-risk-management approaches to incorporate both operational risk and the malicious use of AI, integrating technical and governance controls to mitigate impacts on business and compliance.

Execution plan in three phases

To execute this roadmap with guarantees, we propose the following plan divided into three high-level phases:

PHASE 1 | ACTIVATION OF AI RISK

Activate governance and visibility of AI risk.

- Cross-functional team + clear leadership.
- Operating model (monitoring and war rooms).
- Initial assessment + AI risks + third parties.

PHASE 2 | BUILDING THE AI PROGRAMME

Design, govern and activate the AI security programme.

- Programme defined with clear timeframes and scope.
- Tracking office + indicators.
- War rooms for critical IT changes.

PHASE 3 | VALIDATION AND DEPLOYMENT

Align, validate and activate the programme with stakeholders.

- Communication to CxO / Board.
- Impact on IT processes.
- Response plan (regulators and clients).
- AI risk committee and final approval.

PHASE 1

Risk awareness and team mobilisation:

In this first phase, it is necessary to drive an awareness of the risk associated with artificial intelligence and to mobilise the organisation by creating a cross-functional working group that includes the key areas of security, technology, risk and compliance. This team must operate under a clear organisational model, with continuous-monitoring mechanisms, intensive coordination spaces and defined owners. Without clear leadership and a governance model activated from the very first moment, the organisation responds to offensive AI in a reactive and fragmented way, which is precisely the posture that the adversary exploits.

In parallel, a state-of-the-art analysis of the current situation must be carried out to determine the real level of existing protection, as well as to define the information and evidence to be requested from third parties regarding security. Finally, it is essential to formalise a specific register of AI-associated risks and integrate its management within the control and oversight mechanisms of the second line of defence, ensuring its alignment with the organisation's global risk-management model.

PHASE 2

Building the programme and the monitoring plan:

In this second phase, the construction of the security programme and the definition of a structured monitoring plan are addressed, adapted to the needs and capabilities of the organisation. To this end, an action programme with defined, progressive timeframes must be developed, organised into clear time horizons (for example, 0 to 45 days and 45 to 90 days) that allow for an agile, controlled implementation.

Likewise, it is necessary to establish a programme tracking office, responsible for coordinating initiatives and ensuring their progress through a model of control indicators that facilitates the measurement of progress and decision-making. Complementarily, specific intensive-coordination spaces must be enabled for those initiatives that involve significant changes to traditional IT models and processes, thereby ensuring a controlled transition aligned with the organisation's strategic objectives.

- **Develop a security programme adapted to the entity with bounded, reduced timeframes:**
  - **Contain:** prioritise the immediate patching of critical vulnerabilities, prevent deployments with significant flaws, and reinforce development with AI-based automated detection. Complete this approach with provider alignment and contingency plans adapted to AI-based attacks, ensuring operational readiness against new threats.
  - **Build:** implement continuous exposure management, integrate security into development with AI, and reinforce basic controls (Zero Trust, least privilege and segmentation). Complement with updated third-party management, advanced detection supported by threat intelligence, and protection of attack-resistant backups
  - **Mature:** implement continuous AI-based security testing, reinforce data governance in AI flows, and evolve controls towards advanced architectures (Zero Trust, identity-threat detection and browser isolation). Incorporate autonomous-operation capabilities in the security centre, predictive attack analysis and automated response at machine speed, supported by procedures adapted to AI-based threats. All of this under specific reference frameworks and driven by specialised roles that lead the adoption of AI-based cybersecurity.
  - **Horizon:** implement a continuous operation for vulnerability management and internal AI-based testing, eliminate risk through architectures capable of anticipating and self-recovering. Incorporate advanced models such as digital twins

and secure agentic operations, with dynamic identities and on-demand access. Reinforce the supply chain with early-warning capabilities and enable AI-assisted automatic recovery, aligned with operational-resilience requirements.

- Creation of a programme tracking office and development of a model of monitoring indicators.
- Specific working group for initiatives that affect traditional IT models and procedures.

PHASE 3

Alignment, validation and launch:

This third phase addresses the alignment, validation and launch of the programme with the organisation's main stakeholders. To this end, it is necessary to present an executive summary of the programme and its evolution, aimed at senior management and the board, ensuring their understanding and support. Likewise, the expected impact on IT processes must be clearly detailed, especially in those areas that require transformation or adaptation.

Complementarily, it is fundamental to prepare a coordinated response plan for possible requirements from regulators and clients, ensuring coherence and transparency. In addition, a specific AI risk committee must be assessed and, where appropriate, established, or this matter integrated into the existing risk committee, ensuring adequate governance. Finally, this phase culminates with the formal approval of the programme and the start of its operational execution, consolidating its deployment in the organisation.

4.7. Ten key recommendations

1 Self-assess the level of protection

Define a continuous evaluation model for AI-specific risk, integrating detection, response and governance capabilities. With the form available in INES you can assess your current AI-specific risk level.

2 Create an AI culture

Strengthen organisational knowledge around the impact of AI on cybersecurity, ensuring that those responsible understand both its risks and its operational implications.

3 Establish AI governance as a basic pillar

Implement a governance model that makes it possible to control the use, risks and capabilities of AI within the organisation in a structured way.

4 Reduce the patching gap

Assume that the window between the publication of a patch and its exploitation is minimal and constitutes a real operational risk. Prioritise the identified critical assets and the potentially exploitable vulnerabilities.

5 Prepare for a massive increase in reported vulnerabilities

Size the vulnerability-management team or service to scale up by an order of magnitude.

6 Detect vulnerabilities before deploying

Assume that any vulnerability in production will be discovered first by an attacker.

7 Continuously review code already deployed

Identify and correct existing vulnerabilities in code that is already in production.

8 Design systems assuming compromise

Adopt a design approach based on the premise that the system will be compromised.

9 Actively manage the attack surface

Ensure the visibility and continuous control of all exposed assets. Administration bodies can do so with ELSA.

10 Reduce incident-response times

Adapt response capabilities to the pace of current threats. Given the competitive advantage that AI represents from an offensive standpoint, recovery and resilience capacity is more critical than ever.

5. ANNEX A. Glossary

TERM / ACRONYM	MEANING
AI Red Team	Team or offensive-testing capability specialising in AI systems, models, agents and their ecosystem.
AI RMF	AI Risk Management Framework. NIST's AI risk-management framework.
AI-SPM	AI Security Posture Management. Management of the security posture of AI systems and services.
API	Application Programming Interface.
ATLAS	Adversarial Threat Landscape for Artificial-Intelligence Systems. MITRE knowledge base on adversarial tactics and techniques against AI systems.
CISA	Cybersecurity and Infrastructure Security Agency. The US cybersecurity and infrastructure-security agency.
CNAPP	Cloud-Native Application Protection Platform. Integrated protection platform for cloud-native applications.
CSA	Cloud Security Alliance. Reference organisation in cloud security.
CVE	Common Vulnerabilities and Exposures. Standardised identifier of known vulnerabilities.
CVSS	Common Vulnerability Scoring System. System for scoring the severity of vulnerabilities.
Deepfake	Synthetic content generated or manipulated with AI to imitate the voice, image or video of a real person.
ENS	National Security Framework (Esquema Nacional de Seguridad).

5. ANNEX A. Glossary

TERM / ACRONYM	MEANING
EPSS	Exploit Prediction Scoring System. Predictive system that estimates the probability of a vulnerability being exploited in the next 30 days.
FIDO2	Fast Identity Online 2. Standard for strong, phishing-resistant authentication.
AI	Artificial intelligence.
Agentic AI	An AI system capable of perceiving its environment, making decisions and carrying out actions autonomously to achieve goals, with minimal human intervention.
Generative AI	An AI system capable of producing new content —text, image, audio or code— from patterns learned during its training.
Offensive AI	The use of AI systems to plan, accelerate or carry out cyberattacks, increasing their speed, scale and sophistication and lowering the attacker's barrier to entry.
ISO/IEC 42001	International standard for artificial-intelligence management systems.
IT	Information Technology.
JIT	Just-In-Time. Granting access or privileges only when needed and for a limited time.
KEV	Known Exploited Vulnerabilities. Catalogue of known, actively exploited vulnerabilities maintained by CISA.
LLM	Large Language Model.
Logging	Recording of events, actions or transactions for auditing, monitoring and investigation.
MFA	Multi-Factor Authentication.
MRM	Model Risk Management.
NIST	National Institute of Standards and Technology. The US standardisation and technical-reference body.
OWASP	Open Worldwide Application Security Project. Global community and reference in application security.
OWASP ASI	OWASP Agentic Security Initiative. OWASP initiative focused on the security risks and controls for applications and autonomous agents.
Passkeys	Modern credentials based on asymmetric cryptography that replace traditional passwords.

5. ANNEX A. Glossary

TERM / ACRONYM	MEANING
Pipeline	Automated chain of development, validation, testing and deployment processes.
Prompt injection	An attack that manipulates instructions or context to alter the intended behaviour of a model or agent.
Sandboxing	Isolated execution of processes or code in a controlled environment to limit impact.
SAST	Static Application Security Testing. Static security analysis of source code.
SBOM	Software Bill of Materials. Structured inventory of an application's software components.
Shadow AI	Use of AI tools or services outside the organisation's official governance, inventory or control.
SIEM	Security Information and Event Management. Platform for the management and correlation of security events and information.
SMS - MFA	Multi-factor authentication based on SMS messages.
SOC	Security Operations Center.
SOAR	Security Orchestration, Automation and Response.
SSPM	SaaS Security Posture Management. Management of the security posture in SaaS applications.
Threat intelligence	Analysed information about actors, campaigns, techniques, vulnerabilities and indicators of compromise.
TLPT	Threat-Led Penetration Testing. Advanced intrusion testing guided by threat intelligence and based on realistic adversary tactics.
War Room	Space or dynamic of intensive coordination for managing crises, incidents or critical initiatives.
XDR	Extended Detection and Response. Detection and response extended across multiple security layers.
Zero Trust / ZT	A security model that does not trust any user, device or flow by default and requires continuous verification.

centro criptológico nacional

Security Recommendations to Counter the Offensive AI Models

BEST PRACTICE GUIDE

www.ccn.cni.es

www.ccn-cert.cni.es

oc.ccn.cni.es