
AI-ERA VULNERABILITY MANAGEMENT

*A working paper on Vulnerability Management in an AI-era,
drawn from the AI Security Forum 2026*

TABLE OF CONTENTS

Executive Summary 3

A Dual Vulnerability Response Strategy 5

Vulnerability Reporting..... 6

Building AI Security Ecosystem Collaboration..... 7

AI Augmented Guardrails 9

Generative Application Firewalls..... 11

Looking forward to AI-Era Vulnerability Management 13

Conclusions 16

Further Reading..... 18

EXECUTIVE SUMMARY

The history of vulnerability management can be understood as a progression through distinct eras: from passive defence, to compliance-led scanning, to risk-based management, and now to the age of AI agents. Across each stage, the core objective for security suppliers has remained consistent: reducing the harm caused by vulnerabilities and helping customers mitigate risk more effectively.

Five topics from the forum afternoon track highlight the current challenges:

1. **Industrialised attacks.** The cybersecurity landscape is being reshaped by increasingly professional and industrialised attacks. Automated tools are lowering the technical barriers to entry, enabling a wider range of attackers to operate at greater scale and speed.
2. **Expanded attack surface.** At the same time, the attack surface has grown dramatically. It now includes publicly accessible APIs, multi-cloud environments, SaaS platforms, shadow IT providers, AI-native applications and, increasingly, autonomous AI-based agents. AI-native systems introduce additional threats, including model poisoning and adversarial inputs. Together, these changes are driving a profound transformation in the technical, organisational and international structures of cybersecurity.
3. **AI in offensive and defensive security.** AI is also changing how security work is conducted. LLMs are increasingly used in red teaming, while cybersecurity teams are using AI to improve productivity and support vulnerability research. These tools can help skilled researchers identify vulnerabilities more quickly, but they can also generate high volumes of false-positive vulnerability hypotheses. As a result, AI-assisted findings require adapted triage techniques and careful filtering by human experts.
4. **Speed is now the defining challenge.** The threat landscape is accelerating rapidly. Vulnerabilities exploited before public disclosure have increased by 42%, while attacks from AI-enabled adversaries have risen by 89% year on year. AI-assisted reverse engineering is compressing response windows from weeks to hours, making timely action essential.
5. **Critical infrastructure faces particular pressure.** For operators of energy, water and core telecom networks, the central challenge is to apply safe, reliable updates to live systems without disrupting business continuity.

Figure 1 below represents this new AI-era vulnerability landscape.

Building on these findings, this paper explores how vulnerability management may need to evolve as AI reshapes both the systems being protected and the capabilities of attackers and defenders. The forum did not aim to define or agree a formal AI-era vulnerability management framework, but its discussions and presentations offer a useful basis for identifying key challenges and possible framework elements. Global Digital Foundation presents this section as an initial contribution to that discussion, intended to encourage further input rather than prescribe a final model. The work can be developed further as stakeholder contributions and industry developments emerge, with the aim of supporting a more complete **AI-era vulnerability management** framework and dedicated follow-up work in 2027.

FIGURE 1

Vulnerability landscape in the AI era

Five topics from the forum afternoon track highlight the current challenges

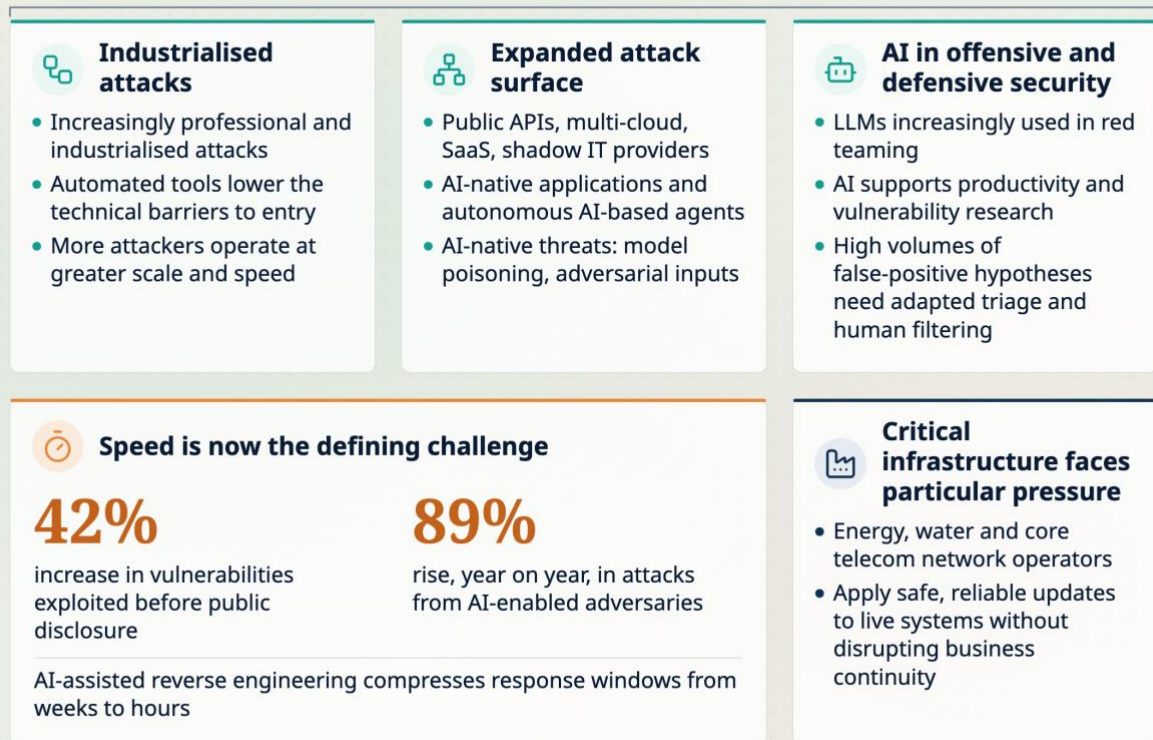


Figure 1 - Vulnerability Landscape in the AI Era

A DUAL VULNERABILITY RESPONSE STRATEGY

Response must combine prevention and reduction. Organisations should transform their vulnerability management practices around two complementary approaches, see Figure 2:

- **Exposure prevention:** strengthening AI and cybersecurity governance to reduce the likelihood of exploitable weaknesses emerging or remaining unmanaged.
- **Exposure reduction:** using automated countermeasures, AI-assisted vulnerability management and, ultimately, agentic exploit detection and response to reduce the impact of active threats.

FIGURE 2

Two complementary approaches

Response must combine prevention and reduction

Organisations transform vulnerability management practices around both approaches

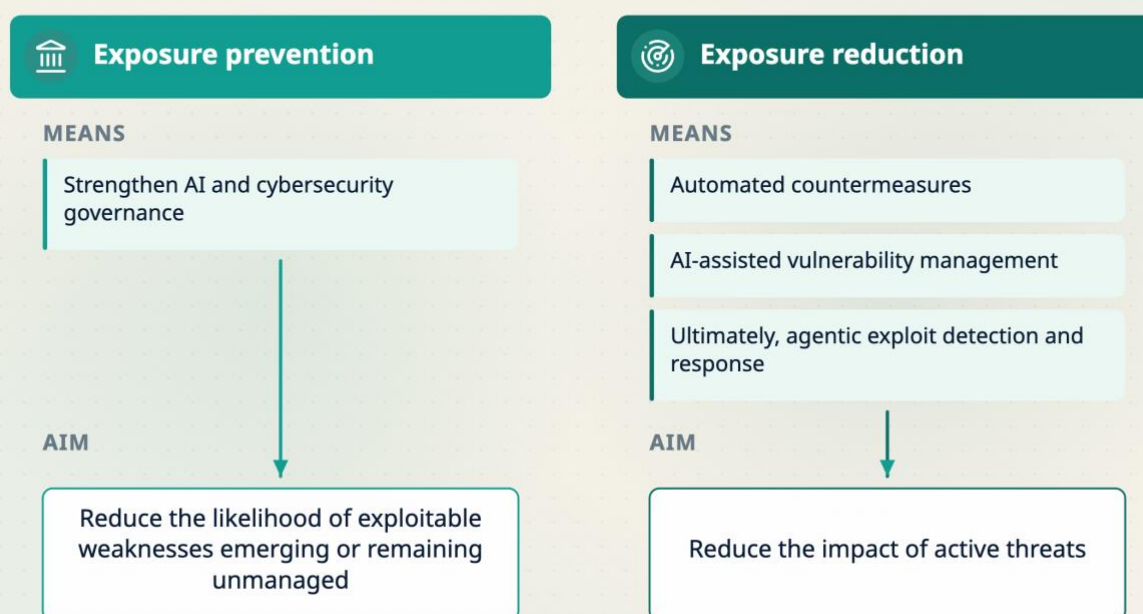


Figure 2 – Response to increased threat through exposure prevention and reduction

VULNERABILITY REPORTING

Managing the growth in vulnerability reports. A major challenge is the rising volume of reports and vulnerability findings submitted by the white-hat community, whose crowdsourced testing provides valuable insight into security weaknesses. Analysing these reports is placing increasing pressure on the platforms that manage bug bounty programmes, where researchers are rewarded when a genuine vulnerability is confirmed.

Using AI for triage. AI is now routinely used by researchers to identify and report potential vulnerabilities, increasing both the overall reporting rate and the number of false positives. In response, platforms are also using AI to triage submissions more efficiently and prioritise human review on the most credible findings. The process is illustrated in Figure 3.

Improving community incentives. Some platforms are also introducing credit-token systems to manage submission quality. Researchers receive a limited stock of submission tokens, each programme sets its own submission cost, and hunters gain or lose tokens depending on the triage outcome.

FIGURE 3

Managing the growth in vulnerability reports

AI-assisted triage and community incentives on bug bounty platforms

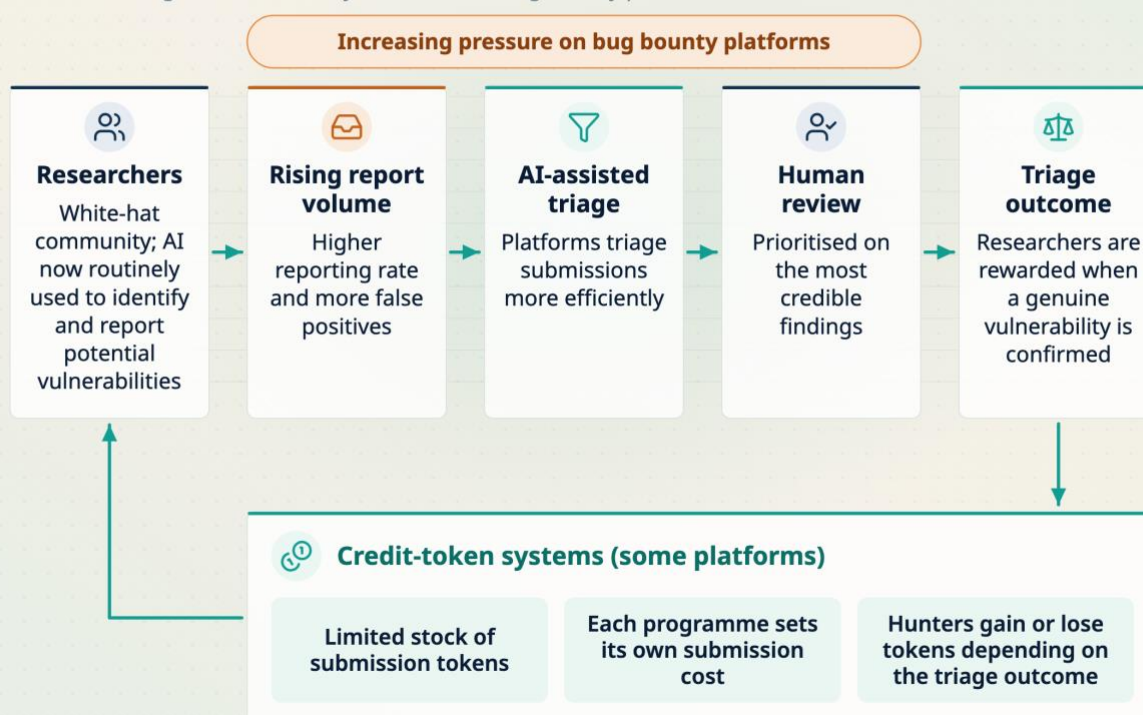


Figure 3 – AI-assisted triage and submission tokens in vulnerability reporting

BUILDING AI SECURITY ECOSYSTEM COLLABORATION

Strengthen vendor–operator collaboration. The AI-enhanced threat landscape requires open collaboration between vendors and operators. This includes co-governance of the ecosystem and joint defence against expanding, complex risks. Practical mechanisms may include regular information sharing, collaborative analysis and remediation processes, coordinated upgrades, and combined security drills. No single entity can defend against AI-driven attacks in isolation. A "Joint Defense" model is required.

Vendors should move towards responsible AI disclosure by providing comprehensive composition management across open-source components, third-party software and AI models, together with immediate mitigation strategies.

Operators should complement vendor advisories with accurate inventories of deployed products and strong local risk-management capabilities.

Focus on deployed operational systems. Collaboration must be grounded in the realities of live systems. System configuration is critical, because vulnerability risk depends on both inherent product vulnerabilities and deployment context such as configuration, exposure, reachable attack paths and compensating controls. As frontier AI systems become more powerful, control, validation and auditability become increasingly important.

Keep accountability with humans. Asset identification, risk assessment, patch approval and compliance evidence remain human responsibilities. AI can inform and accelerate these decisions, but humans remain accountable for the outcome.

Test resilience against adversarial threats. Specialist organisations are beginning to provide facilities for assessing the resilience of AI models against adversarial manipulation, poisoning attacks and model evasion techniques. These assessments examine how well models resist manipulation, alteration or poisoning, and provide important insight into potential weaknesses.

Assessment scenarios should include relevant domains such as 5G and IoT, and should cover AI models, connected systems, cybersecurity defences and humans-in-the-loop. This supports end-to-end vulnerability discovery, assessment, mitigation and validation.

Connect research to practice. Research labs are often one step ahead in the effort to secure AI systems. Explainable AI is one area where research can deliver practical benefits, including explainable network intrusion detection in AI-assisted cybersecurity environments.

Use community standards and shared resources. OWASP¹, the Open Worldwide Application Security Project, is a global, community-led nonprofit focused on improving software security. Founded in 2001, it provides free documentation, tools, standards and forums to help organisations build and maintain trustworthy software. OWASP's *Top 10 for Agentic Applications* identifies important threats for agentic AI applications and highlights novel attack vectors facing LLM-based AI agents. Addressing these risks requires new defensive approaches, such as Generative Application Firewalls.

Build the ecosystem together. Vendors, operators, the open-source community, third-party security providers and research labs must work together to build a secure international AI ecosystem, see Figure 4 below.

¹ OWASP - <https://owasp.org/>

FIGURE 4

Building the ecosystem together

Vendor-operator collaboration, supported by community, specialist and research contributions

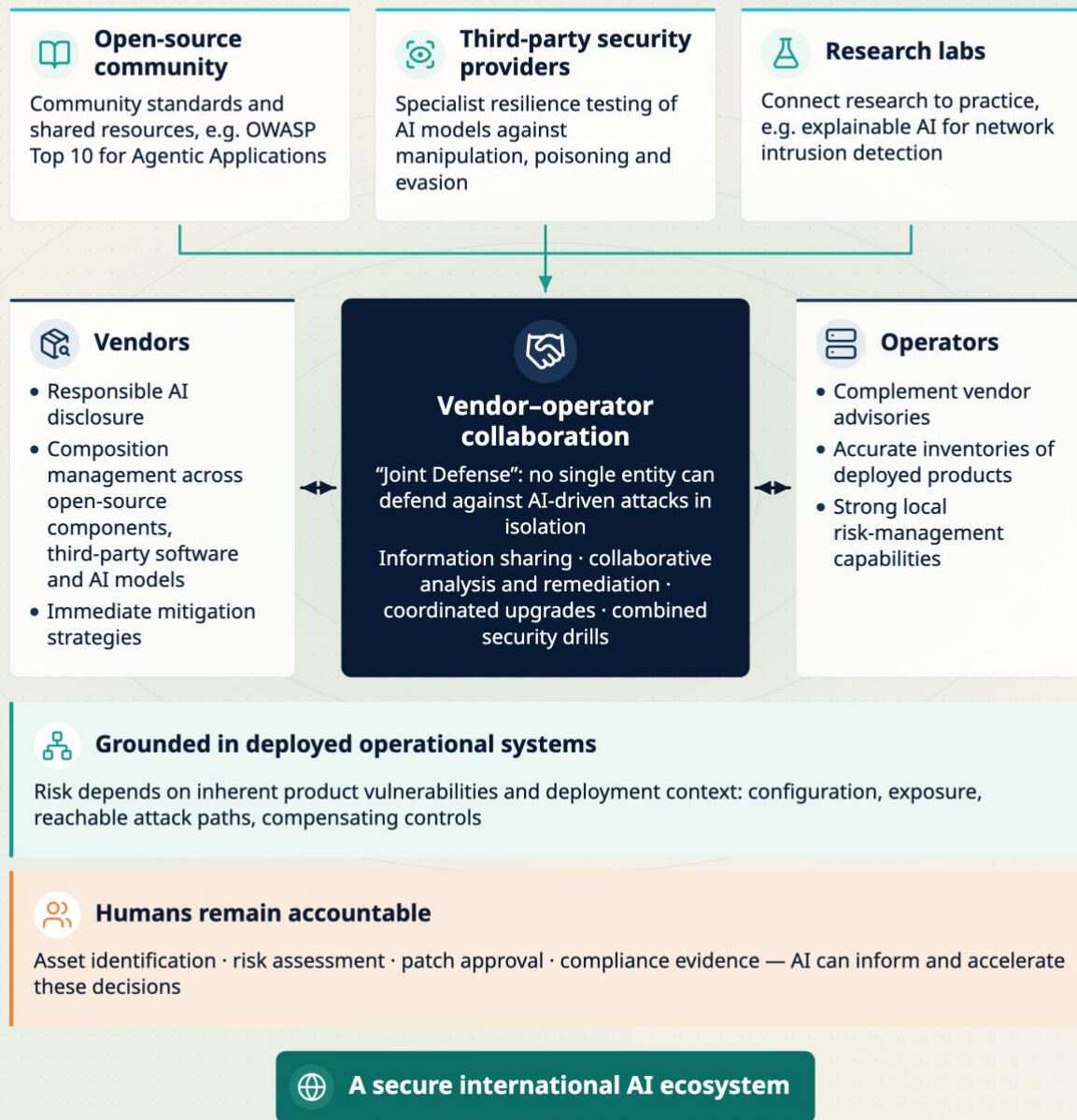


Figure 4 – AI Security Ecosystem Collaboration

AI AUGMENTED GUARDRAILS

Large language models are inherently probabilistic, so their behaviour cannot be guaranteed through prompting alone. As organisations deploy AI assistants and autonomous agents in customer-facing and business-critical settings, they need mechanisms that keep these systems aligned with organisational policies, industry regulations and defined risk boundaries.

Key risks without guardrails. Without appropriate controls, AI systems can hallucinate, provide inappropriate or unauthorised advice, expose sensitive information, become vulnerable to prompt injection and jailbreak attacks, or behave in ways that create legal, regulatory and reputational risk.

Greater pressure in regulated sectors. These challenges are especially significant in highly regulated sectors such as financial services and healthcare. In these environments, organisations must demonstrate that AI systems operate safely, reliably and in accordance with requirements such as the EU AI Act², GDPR³, CRA⁴ and DORA⁵.

Governance must scale with enterprise AI. Existing governance processes often struggle to keep pace with the scale and complexity of enterprise AI deployments. Organisations increasingly operate multiple foundation models, AI agents and retrieval-augmented generation applications across cloud providers and on-premise environments.

Continuous assurance is essential. Traditional governance approaches based on documentation, periodic audits and manual reviews provide limited visibility into day-to-day system behaviour. They also offer little assurance that policy violations will be detected before they have operational consequences.

Required capabilities. As AI becomes embedded in business processes, organisations need continuous monitoring, automated policy enforcement and comprehensive audit trails. These capabilities provide evidence of compliance while enabling AI to be deployed safely at enterprise scale.

Industry must move beyond **Static Vulnerability Management**, focused on tracking CVEs, toward **Dynamic Exploitability Analysis**, focused on understanding real attack paths. This requires:

- **Expanded Asset Intelligence:** Extend visibility beyond software components to include the lineage of datasets, AI models and skill sets, and integrate this intelligence into vulnerability workflows.
- **End-to-End Automation:** Replace human-led workflows with AI-orchestrated processes where appropriate.

Separate security from generation. AI guardrails perform a different role from large language models. Rather than generating content, they act as independent evaluators that sit alongside the primary LLM and assess each interaction before it proceeds, see Figure 5.

Evaluate prompts and responses for risk. These separate models can examine prompts and outputs for threats such as prompt injection, data exfiltration, hallucinations, policy violations and unsafe responses. This allows organisations to identify and block risky interactions before they affect users or downstream systems.

Improve consistency and control. By separating the security function from the generative function, guardrails can remain more focused, consistent and easier to govern than the primary model they protect.

² EU AI Act - <https://artificialintelligenceact.eu/ai-act-explorer/>

³ GDPR - <https://gdpr.eu/what-is-gdpr/>

⁴ CRA - <https://digital-strategy.ec.europa.eu/en/policies/cyber-resilience-act>

⁵ DORA - https://www.eiopa.europa.eu/digital-operational-resilience-act-dora_en

Use fit-for-purpose models. While current off-the-shelf LLMs offer strong general-purpose performance, some AI security applications require more specialised approaches. For example, effective and robust guardrails for AI-native applications often depend on highly tuned models. Small language models can be more efficient in these contexts, requiring less processing power and delivering faster responses. Robustness is strengthened by combining specialist safety controls, purpose-tuned small language models and human-in-the-loop validation. Recognised standards and governance frameworks are essential for assessing and managing these alternative technical approaches, especially when they are provided by third-party suppliers.

Adopt layered, real-time defence. Mature governance should support a layered defence strategy that works in real time. Exposure reduction depends on detection, prioritisation and rapid patching, while complementary countermeasures can help disrupt, isolate and identify intruders.

Strengthen standards and governance. International standards remain vital for spreading best practice and supporting regulatory compliance. However, they must continue to evolve to provide clearer normative requirements for the secure and responsible use of AI.

Maintain human accountability. AI tools can support assurance and governance, but they do not currently replace formal compliance obligations. Frameworks and regulations such as NIS2, DORA, ISO 27001 and the EU AI Act still require structured evidence, oversight and accountable decision-making.

FIGURE 5

Separate security from generation

Guardrails assess each interaction before it proceeds

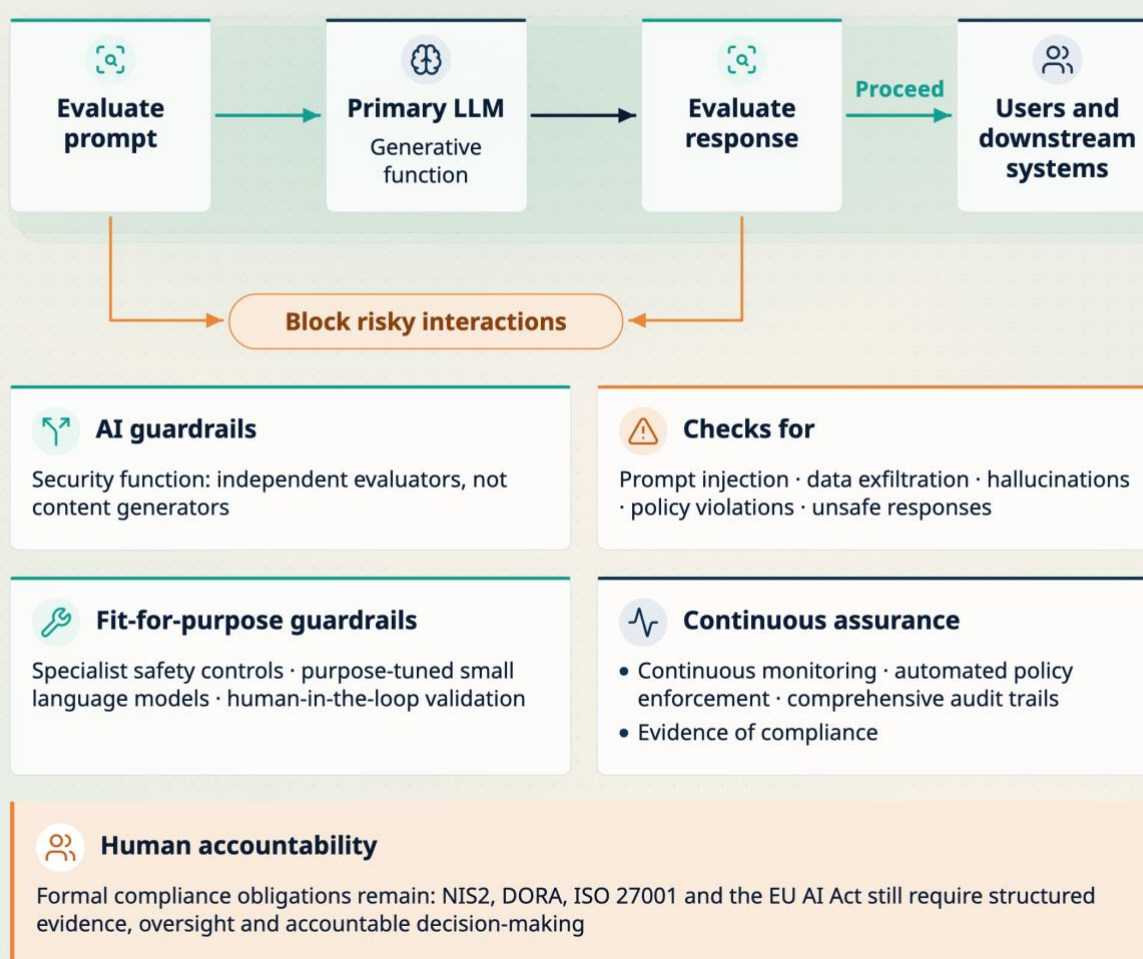


Figure 5 – Guardrails as independent evaluators alongside the primary LLM

GENERATIVE APPLICATION FIREWALLS

A new class of AI security risk. The rapid adoption of generative AI, retrieval-augmented generation (RAG) and increasingly autonomous AI agents is creating cybersecurity challenges that traditional application security tools may not adequately address. Conventional controls such as web application firewalls, API gateways and endpoint protection platforms remain important for AI systems, but may not fully address AI specific risks.

Emerging threats. Organisations must now manage risks that include prompt injection, jailbreaks, sensitive data leakage, malicious tool invocation and manipulation of AI agents through multi-stage interactions.

Governance and lifecycle control. At the same time, organisations need visibility, governance and regulatory compliance across multiple AI models and providers. They must also ensure that AI applications remain secure, trustworthy and resilient throughout their operational lifecycle.

A continuous AI security control layer. Addressing these challenges requires a dedicated AI security architecture that operates between AI applications and the underlying models. Instead of relying only on perimeter defences or periodic vulnerability assessments, this layer should continuously inspect prompts, responses and agent interactions while enforcing policy, monitoring behaviour and detecting threats automatically, see Figure 6.

FIGURE 6

A continuous AI security control layer

Operating between AI applications and the underlying models

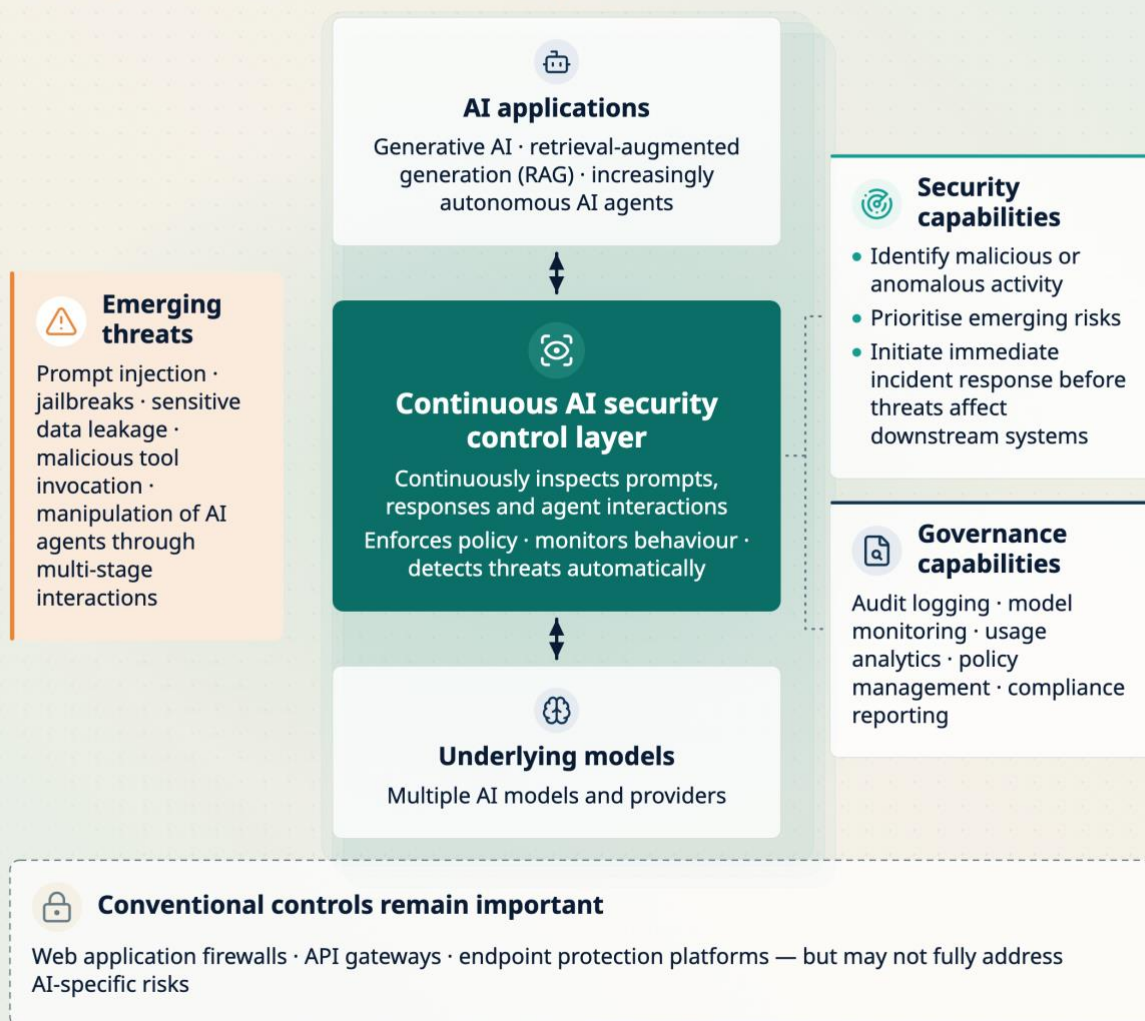


Figure 6 – A continuous AI security control layer between AI applications and models

Security and governance capabilities. AI-assisted analysis helps organisations identify malicious or anomalous activity, prioritise emerging risks and initiate immediate incident response activities before threats affect downstream systems. The same control layer can also support governance through audit logging, model monitoring, usage analytics, policy management and compliance reporting.

Together, these capabilities provide a unified framework for securing AI-enabled applications, reducing operational risk and enabling the safe, responsible deployment of increasingly autonomous AI systems.

LOOKING FORWARD TO AI-ERA VULNERABILITY MANAGEMENT

Security engineering, accurate vulnerability assessment, coordinated disclosure and effective remediation remain the foundations of vulnerability management. AI changes the context by adding new dependencies to AI-enabled products and increasing the scale at which AI-assisted research can find and investigate vulnerabilities. AI can also strengthen proactive vulnerability discovery by combining specialised domain knowledge with established security research and testing techniques, helping identify high-risk and complex logic flaws earlier.

Understanding whether a product is affected may require information about models, datasets, agent skills, dynamically loaded tools and externally provided services, alongside its software components. Component inventories and vulnerability identifiers remain important, while formats such as OASIS CSAF⁶ support the exchange of affected product and remediation information. In determining affect it remains important in the AI Era to understand how a dependency is used and whether the conditions for exploitation are present.

Several standards in the domains of vulnerability management and secure engineering exist. Existing standards include ISO/IEC 30111⁷ covering vulnerability handling⁷, and ISO/IEC 29147⁸ addressing disclosure and communication with users.⁸ Specifically for AI systems, ETSI EN 304 223⁹ sets baseline security requirements for AI models and systems across their lifecycle, including their development, deployment and maintenance. Applying existing standards to AI related findings may require more specific guidance, for example on the evidence needed for validation, the allocation of responsibilities and the criteria for considering a vulnerability resolved.

Bug bounty programs are also adopting AI-assisted techniques: YesWeHack¹⁰ has introduced tools to support vulnerability discovery and help researchers check the evidence and quality of reports before submission. In February 2026, HackerOne¹¹ described a surge of more than 100% in report volume following the release of more advanced AI models and tools. This includes useful discoveries alongside duplicate and unverifiable submissions. Handling larger volumes requires capacity for validation, deduplication and affectedness analysis. Reports concerning model quality, safety or content also need to be distinguished from security vulnerabilities, with coordinated handling where these concerns overlap.

An agent's permissions, access to sensitive data and ability to invoke tools can impact the consequences of a vulnerability affecting an AI system. The Common Vulnerability Scoring System (CVSS)¹² already allows severity assessments to incorporate threat and environmental information. Additional guidance may help practitioners represent exploitability and impact consistently in AI-related cases.

While the number of discovered vulnerabilities has significantly increased year-on-year under the impulse of AI - illustrated by the vulnerability records in the CVE Program¹³ - statistics also show that the proportion of exploited vulnerabilities remains relatively constant (see CISA KEV¹⁴). This points towards attackers continuing to focus on vulnerabilities they find the most useful. However, for vulnerabilities with a public proof of concept,

⁶ OASIS, [Common Security Advisory Framework Version 2.0](#)

⁷ ISO/IEC, [30111:2019: Vulnerability handling processes](#)

⁸ ISO/IEC, [29147:2018: Vulnerability disclosure](#).

⁹ ETSI EN 304 223, [Securing Artificial Intelligence \(SAI\): Baseline Cyber Security Requirements for AI Models and Systems](#)

¹⁰ [YesWeHack Agentic Pentest](#) and [YesWeHack Claude Kit Plugin](#)

¹¹ HackerOne, [Navigating the AI Wave: How We're Keeping Security Research Meaningful](#), 29 May 2026

¹² FIRST, [Common Vulnerability Scoring System Version 4.0](#)

¹³ CVE, <https://www.cve.org>

¹⁴ CISA KEV, <https://www.cisa.gov/known-exploited-vulnerabilities-catalog>

CrowdStrike¹⁵ recorded 88% of the exploitation it observed in January to June 2026 within 48 hours of that proof of concept's release. CrowdStrike also anticipates that AI-assisted discovery and exploit development could compress these timelines further. These data points indicate that the prioritisation of vulnerability handling according to the risk should remain a top priority for industry, together with the shortening and improvement of vulnerability response. AI-assisted investigation could use agent harnesses to combine specialist tools and asset context with knowledge retained from previous assessments. Relevant case histories, validated findings and established assessment methods could then inform subsequent investigations. This helps teams build on earlier analysis and assess exploitability and impact more consistently. AI may also help teams evaluate mitigation options and apply validated protections while permanent remediation is prepared. AI may also support coordinated remediation planning across product and engineering teams, helping allocate remediation effort according to risk, prepare validated fixes and mitigations, and manage deployment across the vulnerability lifecycle.

Telecom, industrial and other critical systems must preserve service continuity and safety while vulnerabilities are addressed. Their testing and deployment requirements often differ from general IT environments, and a mitigation or patch that disrupts an essential service can create additional risk. Automated actions therefore require validation, recovery plans and clear accountability, with human oversight for important decisions. At the same time, expedited deployment of low-disruption mitigations will be essential to reduce vulnerability risk effectively while permanent remediation is prepared and tested.

Vendors remain responsible for identifying and addressing weaknesses in their products, while operators evaluate exposure and response options within their own system environments. Researchers and upstream providers may hold evidence needed by both groups. In the AI era, responsible and coordinated vulnerability disclosure is more important than ever and must adapt to the faster pace created by shorter time-to-exploitation windows. Vulnerability handling should also include a feedback loop into secure engineering, so root causes and lessons from individual findings inform design, development and testing and help prevent recurrence of similar vulnerability classes.

These challenges require stakeholders to work together to advance AI-era vulnerability management, develop industry best practice and standardise key new elements. Priorities should include adapting vulnerability management practices, securing AI-enabled vulnerability management processes, and using tailored AI solutions efficiently and reliably to support vulnerability management decisions, see Figure 7.

Global Digital Foundation intends to continue this work in 2027 with input from Forum participants and other stakeholders. A follow-up paper could explore selected issues in greater depth and set out a fuller AI-era vulnerability management framework, informed by practical experience, further discussion and industry developments.

¹⁵ [CrowdStrike 2026 Threat Hunting Report: Exploitation Window Closes as AI Use Accelerates](#)

FIGURE 7

From affectedness to response

Questions vulnerability handling must answer as AI adds new dependencies

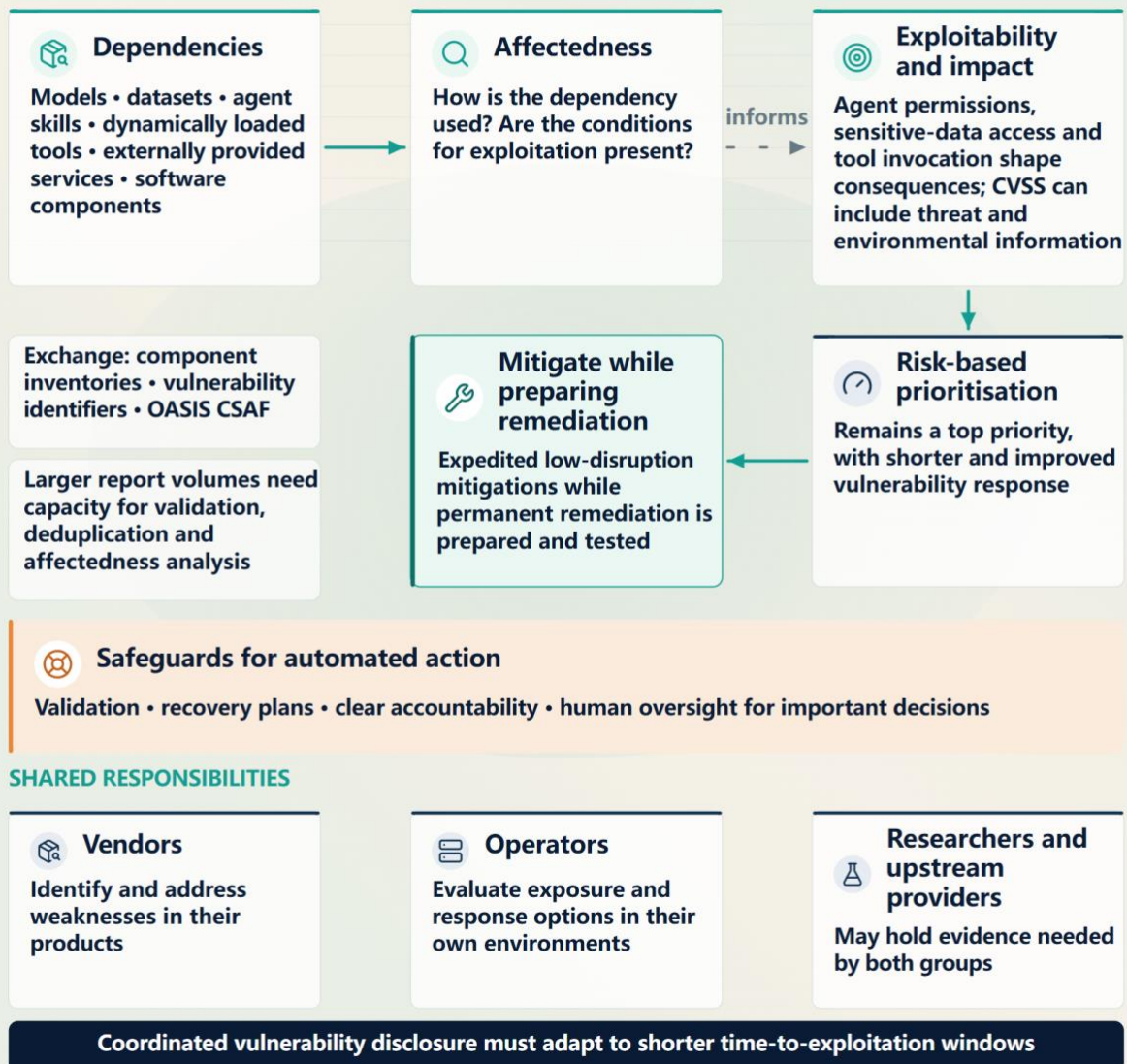


Figure 7 – From affectedness to response, with shared responsibilities

CONCLUSIONS

AI is reshaping vulnerability management by accelerating both offensive and defensive security activity. Industrialised attacks, expanded digital attack surfaces and AI-assisted vulnerability discovery are compressing response times and increasing operational pressure, particularly for critical infrastructure and regulated sectors.

Effective response in this environment requires a dual strategy. Organisations must strengthen **exposure prevention** through governance, standards, assurance and secure design, while improving **exposure reduction** through faster detection, prioritisation, patching and automated countermeasures.

Shortening of the time-to-exploit on certain vulnerabilities: response windows will have to adapt to properly address the vulnerability risk, requiring real-time network awareness and expedited deployment of workarounds. Therefore, vendors should build intelligent patching capabilities, and regularly reserve resources for addressing vulnerabilities to ensure that all vulnerabilities are fixed as soon as possible. Additionally, vendors should establish resilience capabilities such as temporary workarounds for high-risk vulnerabilities, to transition from a passive response approach to an active collaborative defense mechanism.

Four core capabilities are essential for AI defense: proactive intelligence, which uses AI to discover vulnerabilities and strengthen early-warning awareness; autonomous investigation, which automates triage and impact verification so credible risks can be assessed quickly; layered adaptive protection, which applies intelligent shielding across both application and system layers; and autonomous remediation, which uses AI-enabled patching and mitigation to reduce the window of exposure, see Figure 8.

AI-augmented guardrails and generative application firewalls are emerging as important control layers for securing AI-native applications. They provide continuous monitoring, policy enforcement, risk evaluation and auditability across prompts, responses, models and agent interactions.

However, AI does not remove the need for human accountability. Asset identification, risk assessment, patch approval, compliance evidence and final governance decisions must remain under human oversight, supported by clear standards and robust validation processes.

Our assertion is that **AI-era vulnerability management requires coordinated action** across vendors, operators, researchers, regulators, open-source communities and specialist security providers. A secure AI ecosystem will depend on shared standards, collaborative testing, practical assurance mechanisms and sustained investment in both technical controls and organisational governance.

FIGURE 8

Four core capabilities for AI defense

Within a dual strategy, and under human oversight

DUAL STRATEGY



Exposure prevention

Governance · standards · assurance · secure design



Exposure reduction

Faster detection · prioritisation · patching · automated countermeasures

FOUR CORE CAPABILITIES



Proactive intelligence

Uses AI to mine for vulnerabilities and strengthen early-warning awareness



Autonomous investigation

Automates triage and impact verification so credible risks can be assessed quickly



Layered adaptive protection

Applies intelligent shielding across both application and system layers



Autonomous remediation

Uses AI-enabled patching and mitigation to reduce the window of exposure



Human oversight and accountability

Asset identification · risk assessment · patch approval · compliance evidence · final governance decisions
Supported by clear standards and robust validation processes

Figure 8 – Four core capabilities for AI defense under human oversight

FURTHER READING

The following links are suggested for further reading on the topic areas discussed in this report.

Vulnerability Reporting in the AI Era - <https://www.yeswehack.com/product/ai-vulnerability-management>

AI Augmented Guardrails - <https://alinia.ai/guardrails/>

Generative Application Firewalls - <https://neuraltrust.ai/>

Open Worldwide Application Security Project (OWASP) - <https://owasp.org/>