POSITION PAPER

AI-DRIVEN VULNERABILITY DISCOVERY

Executive Summary

AI-driven vulnerability discovery significantly increases the speed, scale and impact of cyberattacks, but it does not invalidate established cyber-security best practices. Instead, it exposes—often within minutes—the weaknesses caused by incomplete, inconsistent or purely procedural implementation of those practices. Success in this new threat landscape depends less on adopting new tools and more on applying proven principles such as segmentation, least privilege, API security, Software Development Life Cycle (SDLC), identity governance, detection engineering and secure-by-design with far greater rigour.

This rigour must be qualitative: best practices need to be enforced as architectural constraints, applied continuously across human, machine and AI identities, and supported by bounded, observable automation. While this disciplined approach cannot eliminate risk, it enables organisations to preserve control, limit systemic impact and remain governable in an environment where both attackers and defenders increasingly operate at machine speed.

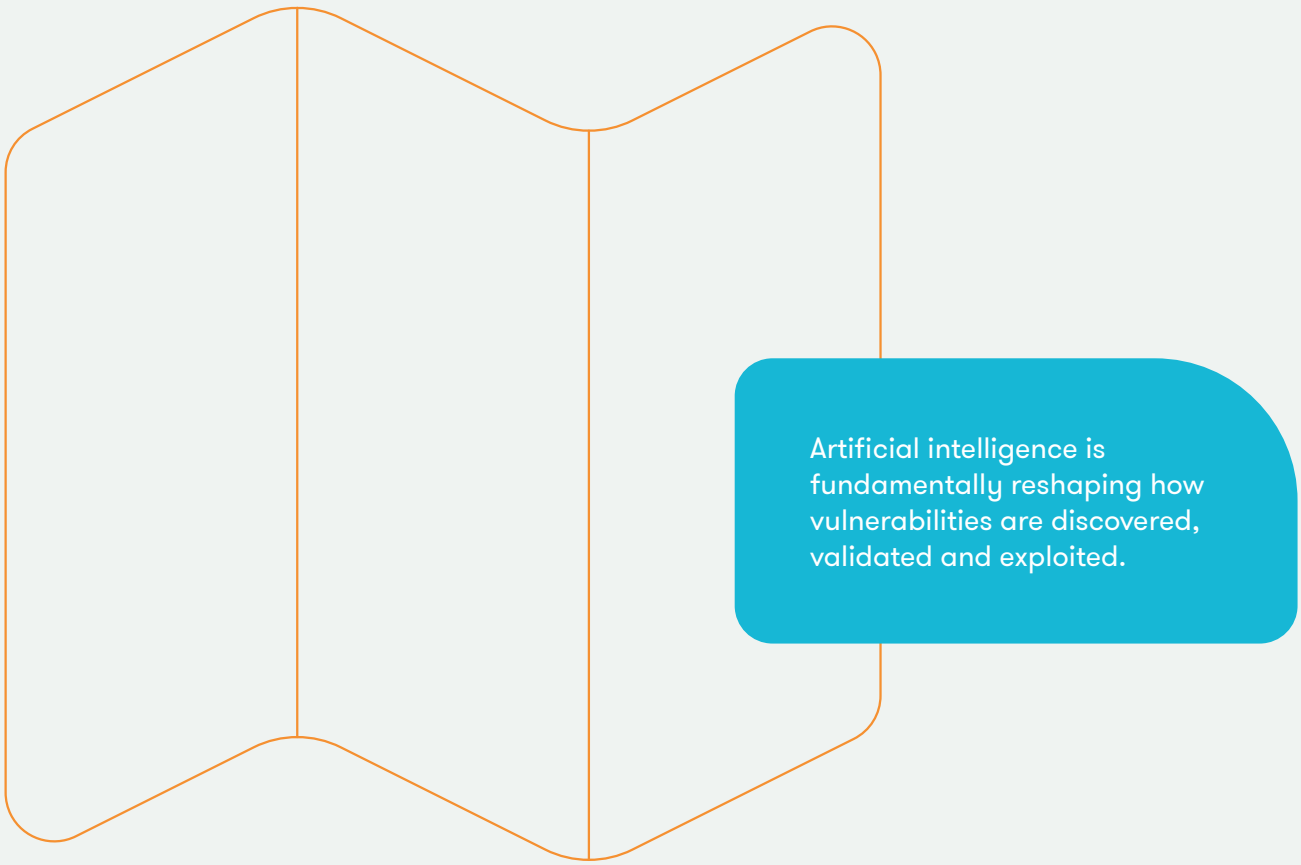
Success in this new threat landscape depends on proven principles and greater qualitative rigour.

01

Defending in the Age of AI-Driven Vulnerability Discovery

Artificial intelligence is fundamentally reshaping how vulnerabilities are discovered, validated and exploited. Recent advances in so-called frontier or agentic AI models mean that tasks once requiring skilled human attackers over weeks or months can now be executed automatically, continuously and at machine speed. This has important implications for how organisations should think about risk, prioritisation and defence.

This article synthesises current industry practice, operational experience and academic research on AI-driven vulnerability discovery, without reference to specific vendors, tools or events.



02

What AI-Driven Vulnerability Discovery Changes

From Assistance to Autonomy

Earlier generations of AI primarily assisted humans: improving code completion, accelerating scans or helping analysts triage alerts. The current shift is towards autonomous agents that:

* Continuously scan environments for weaknesses
* Validate whether vulnerabilities are exploitable in practice
* Chain multiple lower-severity weaknesses together
* Adapt their behaviour without explicit human instruction

The result is not a new type of attack technique, but a dramatic change in scale, speed and persistence.

Speed Is Now the Dominant Risk Factor

Traditional security models assumed a temporal buffer between:

* Vulnerability disclosure
* Exploit development
* Active exploitation

AI-driven discovery collapses this window. Exploitation can begin minutes after discovery, often faster than organisational patching or change-management cycles can realistically operate.

This fundamentally weakens strategies that rely primarily on:

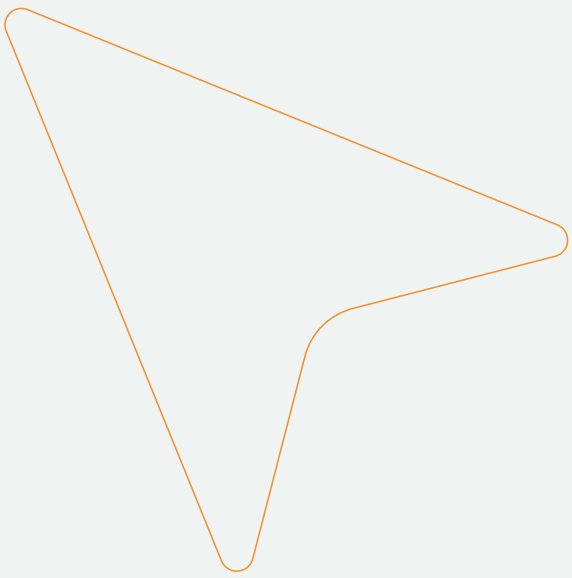
* Periodic scanning
* Patch cycles focused on "critical only" vulnerabilities
* Human-heavy validation workflows

The End of Severity-Only Prioritisation

AI agents do not respect human-centric severity labels. They can efficiently:

* Combine medium or even low-severity issues
* Exploit configuration weaknesses and identity mis-assignments
* Traverse complex dependency and supply-chain paths

As a consequence, an environment with many "acceptable" residual risks can still be fully compromised.



03

Key Risk Domains Introduced or Amplified by AI

Vulnerability Volume and Chaining

AI dramatically increases the volume of discovered weaknesses while making *combinatorial exploitation* practical. Backlog reduction through patching alone becomes mathematically impossible.

Perspective: this is less a patching problem than a risk management problem.

Identity as the Primary Attack Accelerator

Many AI-driven attacks succeed not because of novel vulnerabilities, but because:

* Privileges are excessive or persistent
* Service accounts and agents inherit human authority
* Lateral movement is insufficiently constrained

Identity flaws amplify every other technical weakness.

Supply Chain and Embedded Risk

AI excels at analysing reused components, libraries and legacy code.

Vulnerabilities may exist:

* Deep inside dependencies
* Far from any code actively maintained by the organisation
* In places previously assumed to be "stable" or well-reviewed

AI Systems as New Attack Surfaces

Generative and agentic AI introduce new exposure points:

* Prompt and response manipulation
* Model input poisoning
* Open-source malicious patches
* Unsafe tool-use and function execution
* Shadow AI and unsanctioned agents with data access

These risks sit *alongside* traditional infrastructure and application threats.

Operational Fragility and Automation Risks

As defenders increase automation to keep up, a new tension appears:

* Faster response vs operational stability
* Autonomous remediation vs unintended outages or malware introduction

Blind trust in automation is risky; complete human gatekeeping is too slow.

Perspective: this is less a patching problem than a risk management problem.

04

Defensive Principles That Still Hold and Matter More Than Ever

Despite the novelty of AI-driven attacks, the discussions converged on a clear insight: AI does not break security fundamentals, but it punishes organisations that have treated them as optional or symbolic. Best practices therefore need to be deepened, systematised and executed with far greater discipline.

Reduce Blast Radius, Not Just Vulnerabilities

Given the speed and scale of AI-enabled exploitation, prevention will fail somewhere. The primary defensive objective must therefore be to limit the impact of inevitable failure.

Best practices include:

* Designing for containment through strong network and application segmentation
* Enforcing strict identity boundaries between users, services, workloads and environments
* Applying least-privilege and just-in-time access as defaults, not exceptions
* Ensuring that no single credential, agent or service account can provide end-to-end access to critical assets

In an AI context, small blast radii turn rapid exploitation into a manageable incident rather than a systemic crisis.

Shift from Patch-Centric to Exposure-Centric Risk Management

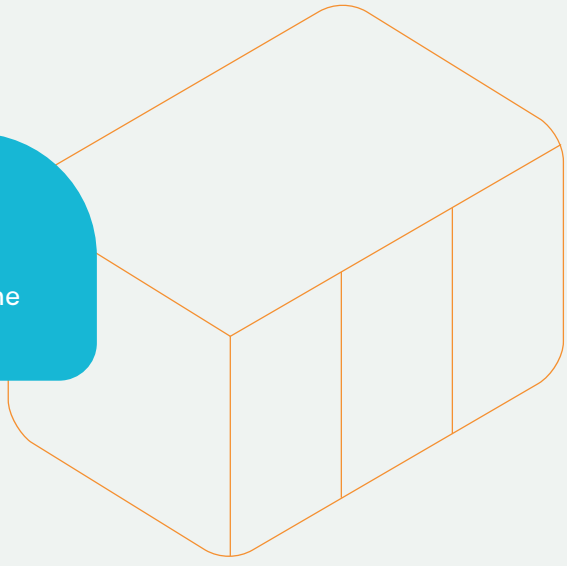
AI breaks traditional vulnerability prioritisation models based solely on CVSS—Common Vulnerability Scoring System—or severity labels. Best practice now requires continuous understanding of real exposure.

This means:

* Knowing where assets are reachable from, not just where they exist
* Understanding how vulnerabilities combine across systems, identities and trust paths
* Factoring in compensating controls and architectural barriers

The governing question should no longer be "*Is this vulnerability severe?*" but "*Does this vulnerability meaningfully increase our current attack paths?*".

Best practices need to be deepened, systematised and executed with far greater discipline



Engineer for Detection and Containment at Machine Speed

Since attackers operate at machine speed, defensive response must also accelerate—without sacrificing judgement.

Best practices include:

* Prioritising signal quality over raw alert volume
* Correlating data across endpoints, identities, networks and applications
* Pre-defining automated or semi-automated containment actions for clearly low-risk scenarios
* Explicitly defining where human approval is required, and where it is not

The goal is not full autonomy, but predictable, bounded automation that reduces mean time to detect and respond.

Treat Identity as Core Security Infrastructure

Across scenarios, identity repeatedly emerged as the most effective attack accelerator—and the most powerful defensive control.

Best practice requires:

* Aggressively reducing standing privileges across users and service accounts
* Applying the same governance standards to machine and AI identities as to humans
* Monitoring identity behaviour, not just authentication success
* Designing identity systems under the assumption that credentials will eventually be abused

In practice, strong identity design often compensates for weaknesses elsewhere.

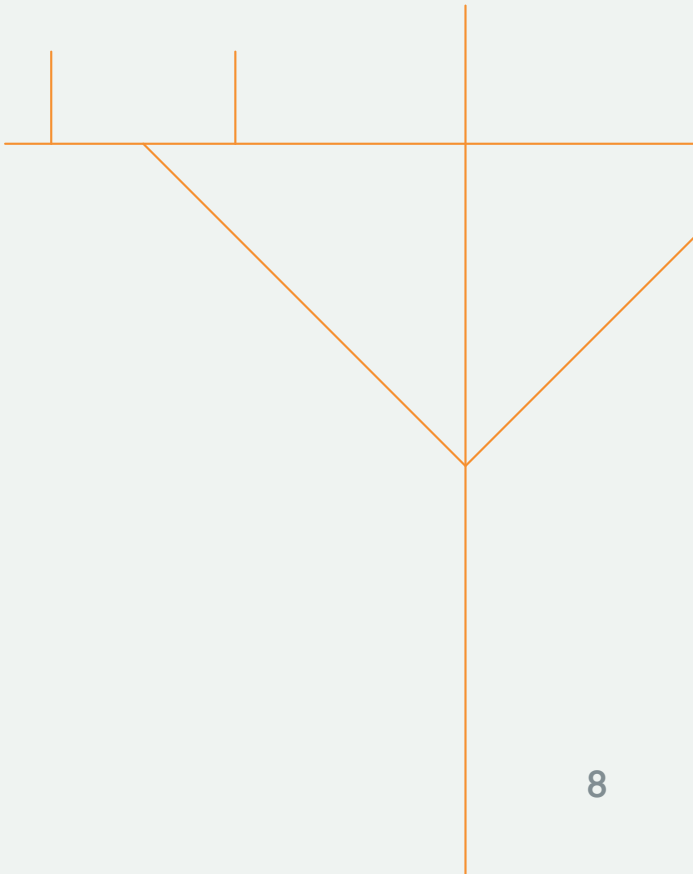
Enforce Secure-by-Design Through Engineering Controls

AI makes secure development more scalable—but only if controls are enforced, not advisory.

Best practices include:

* Integrating security checks directly into development and deployment pipelines
* Continuously analysing dependencies and supply-chain components
* Blocking insecure patterns automatically, rather than documenting them as exceptions
* Using AI to assist developers, while preventing insecure code paths from ever reaching production

Secure-by-design must become an engineering constraint, not a policy aspiration.



05

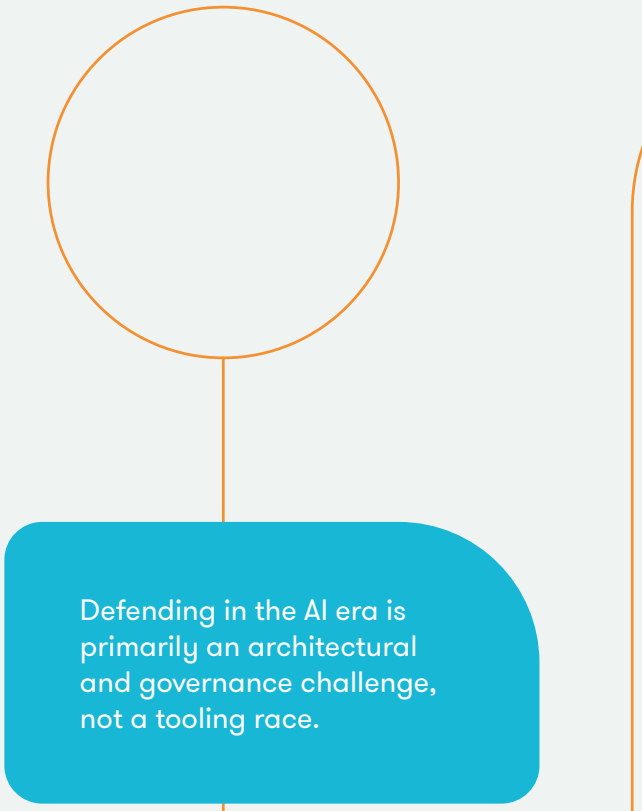
Overarching Best-Practice Conclusions

Taken together, these practices point to a broader conclusion: defending in the AI era is primarily an architectural and governance challenge, not a tooling race.

Across industry practice and academic research, several common principles apply:

* Assume compromise, design for resilience—speed favours attackers, but impact is controllable
* Constrain trust explicitly—human, machine and AI trust relationships must be deliberately limited
* Automate with boundaries, not blind faith—autonomy must be bounded, observable and reversible
* Reduce complexity wherever possible—AI exploits complexity faster than humans can manage it
* Invest for the system you want in five years, while surviving the next two—short-term turbulence is unavoidable, long-term resilience is a choice

Organisations that apply these principles consistently will not eliminate risk. But they will retain control, preserve optionality and remain governable—even as attackers and defenders increasingly delegate execution to machines.



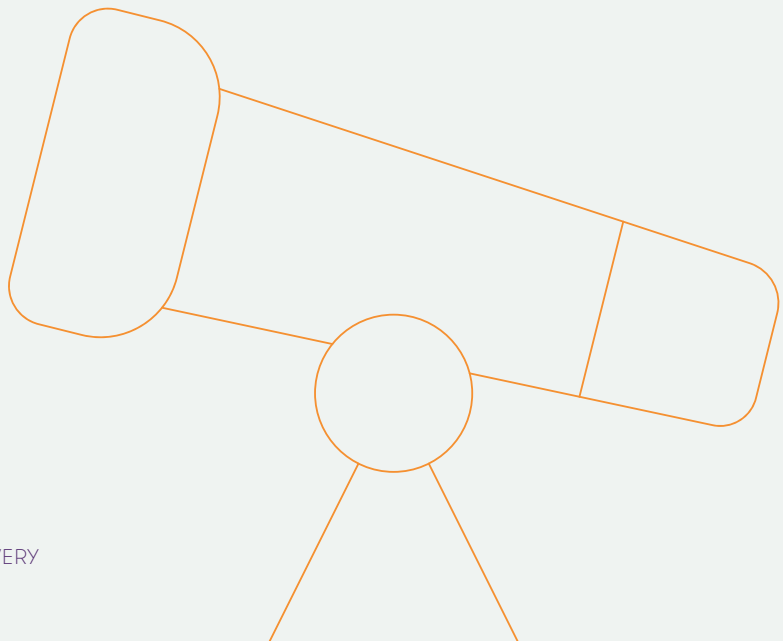
06

Practical Next-Steps: What Every Company Should Do Now

The following practical actions translate the above principles into concrete steps that organisations can apply immediately, regardless of size or sector. They are intentionally vendor-agnostic and focus on operating-model discipline rather than new tooling.

1. **Map and minimise blast radius**
   Document critical assets and crown-jewel paths, then enforce (network) segmentation and access boundaries so that no single account or service can traverse end-to-end.
2. **Kill standing privilege**
   Identify all persistent admin, service and automation accounts; replace them with just-in-time access and time-bound elevation wherever possible.
3. **Reprioritise vulnerabilities by exposure**
   Complement CVSS with reachability and path analysis to focus remediation on vulnerabilities that materially increase real attack paths.
4. **Pre-authorise containment**
   Define a small set of low-risk, automated containment actions (e.g. isolate, revoke tokens, block sessions) that can trigger without human delay.
5. **Govern machine and AI identities**
   Inventory all non-human identities, apply the same lifecycle controls as for users, and monitor behavioural deviations.
6. **Lock security into pipelines**
   Make secure-by-design non-negotiable by enforcing checks in CI/CD and blocking insecure artefacts from reaching production.
7. **Shine light on shadow AI**
   Inventory AI tools and agents with data access, define acceptable use, and block unsanctioned data exfiltration paths.
8. **Simplify before adding controls**
   Remove unused systems, excessive rules and legacy privileges—AI exploits complexity faster than defenders can manage it.

These steps do not eliminate risk, but they significantly increase organisational resilience and buy time in an environment where attackers increasingly operate at machine speed.



07

Academic Perspective: Structural Risks and Long-Term Implications

A complementary and more cautionary perspective comes from academia, articulated by Professor Bart Preneel, a leading cryptographer and long-standing researcher in applied cryptography, software security and systemic cyber risk. His analysis does not contradict industry observations on speed and scale, but places them in a broader context of structural fragility in modern AI systems and digital infrastructures.

AI Systems Are Powerful—and Fundamentally Brittle

A central academic insight is that today's AI systems, including advanced generative and agentic models, are statistical systems whose internal decision-making is not well understood, even by their designers. While they can outperform humans on specific tasks, they can also fail in highly counter-intuitive ways.

Research over more than a decade has shown that:

* Small perturbations or "noise" in inputs can cause radically incorrect outputs
* Guardrails and safety mechanisms can often be bypassed or degraded
* Correct behaviour cannot be reliably inferred from past performance

From a security standpoint, this means that neither offensive nor defensive AI should be assumed to behave predictably under adversarial pressure.

We Are Entering a Destabilisation Phase

According to Prof. Preneel, AI is not merely accelerating existing attack and defence mechanics—it is temporarily destabilising the cyber-security equilibrium. Capabilities that traditionally required expert teams are becoming widely accessible, while defensive architectures and governance models are still adapting.

This creates a period where:

* Attack capabilities scale faster than organisational defences
* Vulnerability discovery outpaces remediation capacity
* Risk exposure grows even in otherwise mature environments

In this phase, optimism about future defensive maturity should not obscure near-term risk.

Speed and Scale take precedence over Human Control

Where industry emphasises speed as an operational challenge, academia highlights its systemic consequences. AI enables exploitation not only to happen faster, but also at a scale where human oversight becomes selective rather than comprehensive.

This leads to difficult trade-offs:

* Full manual control does not scale
* Full autonomy introduces unpredictable failure modes
* Hybrid models require clear boundaries and accountability

The implication is that defensive autonomy must be deliberately constrained, rather than maximised by default.

Guardrails Are Necessary, but Not Sufficient

Academic research strongly supports the use of guardrails, validation layers and policy-driven constraints—but also demonstrates their limitations. Thousands of research papers document how:

* Safety controls reduce risk but rarely eliminate it
* Adversarial probing improves bypass success over time
* Complex systems accumulate unexpected interactions

This reinforces the need for defence-in-depth rather than reliance on any single "AI safety" mechanism.

Expect AI-vs-AI Dynamics

A longer-term insight is that cyber security is moving toward AI systems interacting primarily with other AI systems. Attackers deploy automated agents; defenders respond with automated detection, correlation and response. Over time, human operators increasingly supervise rather than directly intervene.

This dynamic mirrors developments in other domains, such as autonomous systems and modern warfare: escalation is met by counter-automation, not by reverting to manual control.

Long-Term Outlook: Re-architecting for Resilience

Despite near-term pessimism, the academic outlook is cautiously optimistic in the long run. AI also enables:

* Scalable formal verification of software
* Systematic refactoring into memory-safe and verifiable languages
* Hardware-assisted security and isolation mechanisms

What was previously economically infeasible—such as formally verifying large code bases—becomes realistic once AI reduces cost and effort.

However, this path requires fundamental re-architecting of systems, not incremental tooling. It also implies higher upfront investment, clearer regulatory boundaries and explicit societal choices about acceptable risk.

AI-driven vulnerability discovery represents a structural transition rather than a temporary spike.

Key Academic Takeaway

From an academic perspective, AI-driven vulnerability discovery represents a structural transition rather than a temporary spike. The short term will likely be volatile and uncomfortable. The long term can be safer—but only if organisations accept that:

* Complexity must be reduced, not just managed
* Trust needs to be constrained by design
* Automation must be governed as a socio-technical system, not a tool

In other words, resilience will come less from reacting faster than attackers, and more from building systems that remain secure even when automation—on both sides—inevitably fails.

Conclusion

A defensible overarching conclusion from both industry practice and academic research is that the rise of AI-driven vulnerability discovery does not invalidate existing cyber-security best practices—it dramatically raises the bar for how rigorously they must be applied.

The increased speed, scale and autonomy introduced by AI leave little tolerance for incomplete, inconsistent or purely procedural security measures. Practices such as segmentation, least privilege, identity goverance, detection engineering and secure-by-design have long been recognised as effective; AI simply exposes, within minutes or hours, the consequences of treating them as optional, delayed or advisory.

Crucially, this does not mean that organisations can "work harder" within the same operating models. Rigour in the AI era is qualitative, not merely quantitative. It requires best practices to be:

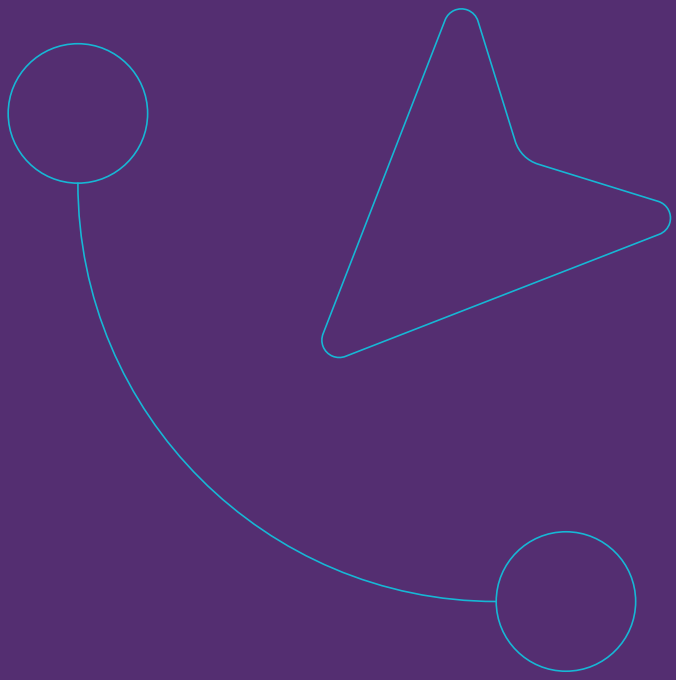
1. Implemented as enforceable architectural constraints rather than policy statements
2. Applied continuously rather than periodically
3. Extended consistently to human, machine and AI identities alike
4. Supported by bounded, observable and reversible automation

The rise of AI-driven vulnerability discovery does not invalidate existing cyber-security best practices—it dramatically raises the bar for how rigorously they must be applied.

From an academic perspective, this rigour does not eliminate risk. AI introduces a phase of structural destabilisation in which perfect defence is unrealistic. However, organisations that apply proven principles with architectural discipline can preserve control, reduce systemic impact and remain governable—even as attackers and defenders increasingly operate at machine speed.

In short, success in this new landscape will belong not to those who chase the latest tools, but to those who finally operationalise well-known principles with the consistency, enforcement and governance that the AI era demands.

Collaboration Across the Ecosystem



Finally, the increasing speed, scale and systemic nature of AI-driven vulnerabilities reinforce the importance of structured collaboration across the cybersecurity ecosystem. No single organisation—public, private or academic—can maintain full situational awareness or resilience in isolation.

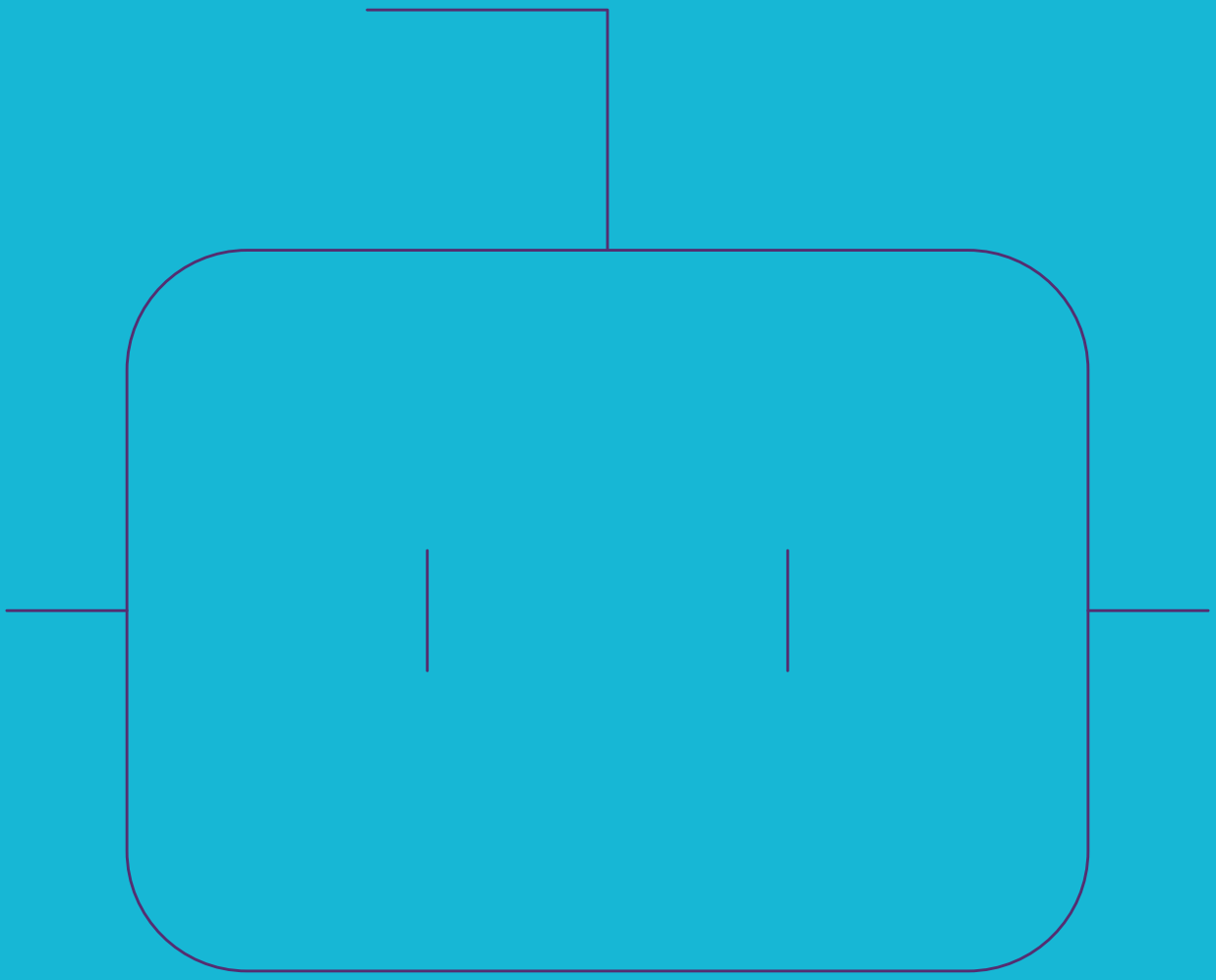
The Cyber Security Coalition actively supports this cross-sector cooperation by facilitating:

1. Sharing of best practices through its working groups and focus communities, ensuring that proven defensive approaches are consistently disseminated and operationalised across sectors
2. Exchange of threat intelligence between members, enabling faster awareness of emerging vulnerabilities, exploitation techniques and attacker behaviour
3. Public–private–academic dialogue, bringing together industry leaders, government bodies and research institutions to align on systemic risks and long-term resilience strategies

Within the Cyber Security Coalition, these activities are embedded in dedicated focus groups and community sessions, where members can continuously exchange insights, validate approaches and translate collective knowledge into actionable improvements.

In an environment where attackers increasingly leverage shared tools, automation and intelligence at scale, defence must evolve as a collective capability—organised, trusted and continuously refined across the entire ecosystem.

Defence must evolve as a collective capability.



Publication date

July 2026

Contact

Cyber Security Coalition

Koning Albert II Laan 4
4 Boulevard Roi Albert II
1000 Brussels

info@cybersecuritycoalition.be
www.cybersecuritycoalition.be

CYBER SECURITY
COALITION