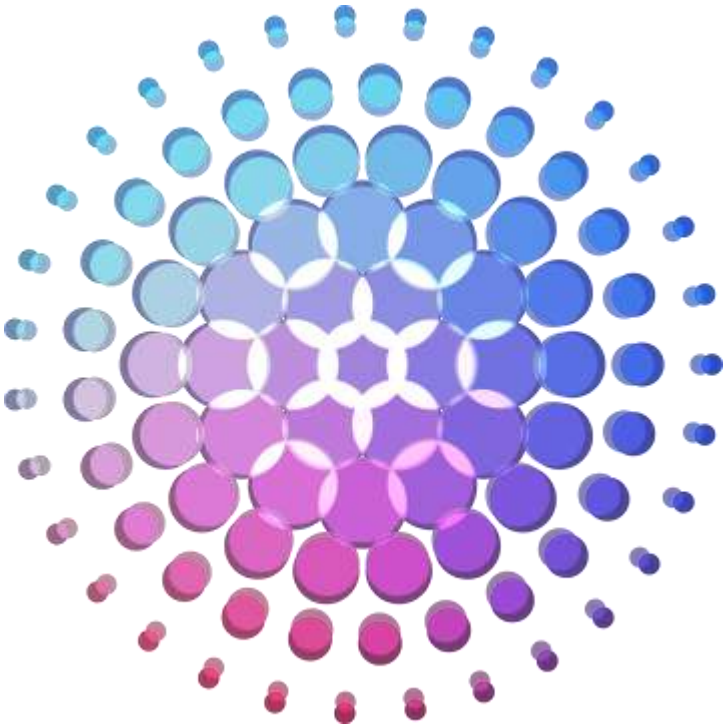
PERSPECTIVE

Preparing for machine speed risk

Frontier AI is accelerating vulnerability discovery and changing the economics of cyber operations. Enterprises should respond by improving their cyber resilience through reducing reachable exposure, strengthening identity, improving detection and containment, governing AI with delegated authority and proving that critical services can recover under attack.



CENTRAL PROPOSITION

The fundamentals have not changed. The urgency has.

Introduction

Artificial intelligence is beginning to reshape both sides of the cybersecurity equation. Enterprises are using AI to improve software development, automate processes, analyse data and support business decisions. At the same time, increasingly capable models can assist with vulnerability research, reconnaissance, exploit development and the orchestration of complex technical tasks.

The significance of this change is not simply that AI may help attackers perform familiar activities more efficiently. It is that advanced models can reduce the time, specialist knowledge and effort required to identify and act on weaknesses. This challenges security operating models built around periodic assessments, monthly patch cycles and human-scale volumes of findings.

Recent cybersecurity benchmarks illustrate the direction of travel. CyberGym evaluates AI agents against 1,507 historical vulnerabilities drawn from 188 open-source software projects. Its wider observatory also evaluates exploit generation and end-to-end workflows covering vulnerability discovery, proof-of-concept development and patch creation. These benchmarks should not be interpreted as direct measures of an attacker's success against a particular enterprise. They do, however, provide evidence that AI capability in vulnerability analysis is advancing quickly.

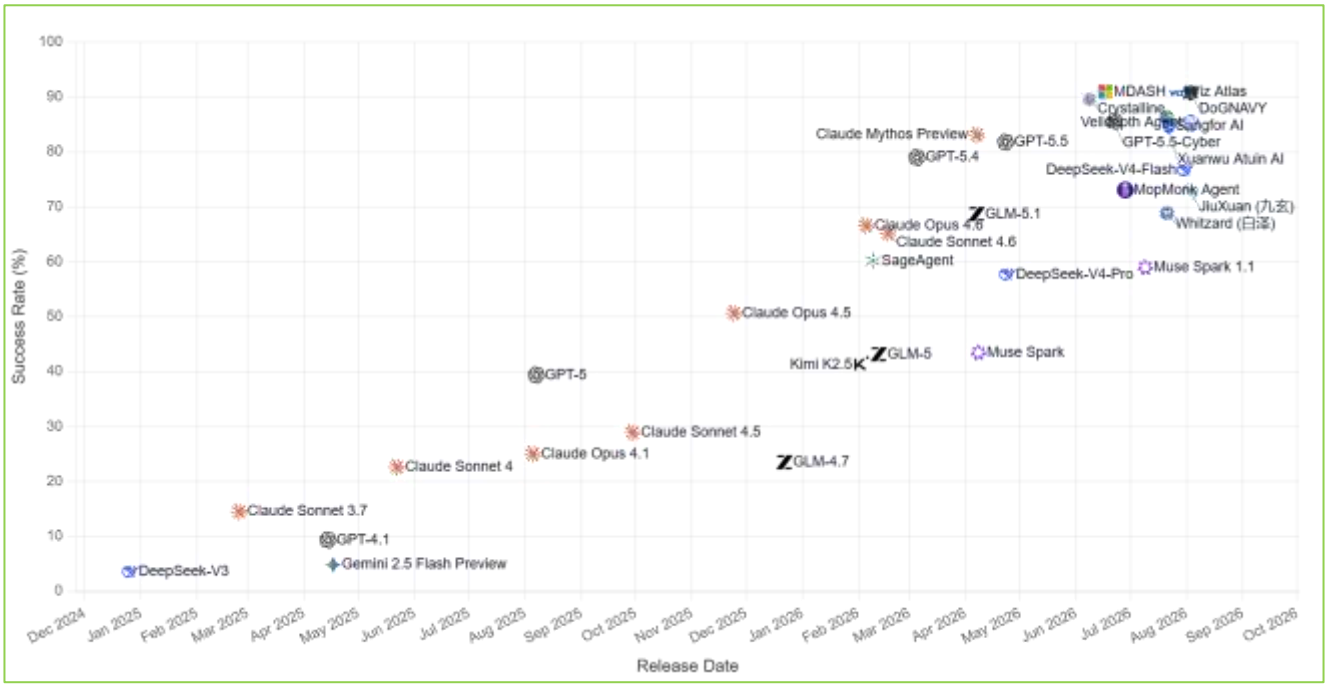


Image 1: The CyberGym leaderboard. Source: https://www.cybergym.io/cybergym/

Recent months have seen an ever-growing mix of models and agents scoring highly on the benchmark. Many of these agents, using specific custom harnesses and leveraging readily available models, perform better than standalone models. This shows that AI models and agents are increasingly getting very good at finding and exploiting vulnerabilities.

Analysis of recent AI-enabled cyberattacks, whether AI-augmented or AI-autonomous, has revealed that the TTPs have not changed, but the speed and scale have. Attackers are primarily using AI for phases of attacks where it is most effective, such as reconnaissance, defence evasion and exploit development, and less for post-exploit phases.

The appropriate response is neither alarm nor complacency. Most attacks will continue to use familiar paths, including compromised credentials, exposed services, phishing, vulnerable applications, cloud misconfigurations and excessive privilege. Attackers generally optimise for cost, reliability and return rather than novelty. Frontier AI becomes particularly important when it helps discover unfamiliar weaknesses, connect several lower-severity issues, interpret custom business logic or shorten the interval between discovery and exploitation.

WHAT HAS CHANGED

Established security practices remain relevant, but the pace and discipline with which they are applied must change.

The exposure boundary is expanding

Traditional attack surface management has largely focused on hosts, applications, networks and externally reachable services. Frontier AI makes a wider set of enterprise context relevant to security.

Source code helps a model understand control paths, authorisation decisions and business logic. Architecture documents, workflow descriptions and API specifications explain how a system is intended to operate. Runtime responses, telemetry and cloud configuration can help establish whether a weakness is reachable and exploitable. Identities, agents, OAuth grants, service accounts and third-party connectors determine how far an actor can move or what actions it can perform.

The resulting question is no longer limited to, “Which assets are exposed?” Enterprises must also ask, “What context and authority could an AI-enabled actor use to discover, validate and act on risk?”

This change is especially relevant for custom applications. A vulnerability in a commercial product may eventually receive a vendor patch. A weakness in proprietary code, such as an authorisation bypass, a tenant isolation failure or a flawed transaction workflow, belongs to the enterprise itself. Such weaknesses may not have a CVE identifier or appear in a conventional vulnerability feed, despite carrying serious business consequences.

AI adoption also creates exposure from within the organisation. Agents and copilots may gain access to mailboxes, repositories, enterprise data, business applications and

automation tools. Shadow AI, excessive connector permissions and weak oversight of agentic systems can create security risks even without an external compromise.

The fundamentals are more important than ever and need a fresh look. For enterprises, the effort must be directed towards raising the cost of attacks for attackers, reducing external and internal exposure, creating friction against misuse and abuse, and finally constraining the blast radius. Tactical responses may offer point gains, but a strategic enterprise-level programme secures the enterprise.

Six connected workstreams for enterprise action

A resilient response requires more than an additional security tool or a standalone AI policy. It requires coordinated action across six workstreams. The value lies in the handoffs: each workstream produces priorities, controls, telemetry or evidence that the others rely upon.

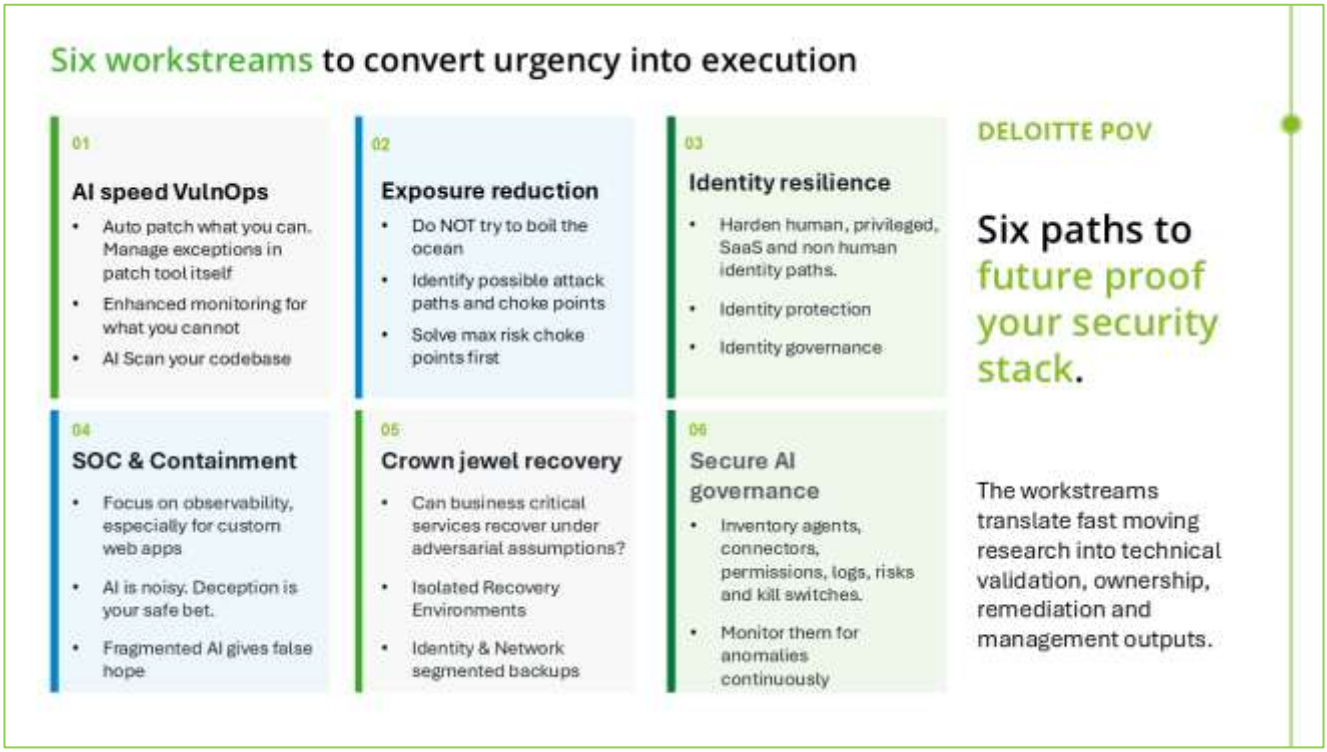


Image 2: The six workstream programme for a resilient enterprise

1. AI speed vulnerability operations

Traditional vulnerability and patch management programmes follow a cycle that could last months between a vulnerability discovery and patching. This is incompatible with modern truth, where AI can discover and exploit known or custom vulnerabilities within minutes or hours. Vulnerability management should become a continuous operational capability. Enterprises should establish differentiated remediation service levels, increase automation, strengthen exception governance and connect vulnerability decisions to asset context and business impact.

Software composition visibility is also important. Security teams should understand which open source and third-party components support critical services and where those components are deployed. For custom applications, code analysis, testing and remediation should be integrated into software engineering rather than treated as a periodic assurance exercise.

The critical bottleneck is moving from finding issues to making decisions. Security teams will need the authority, data, and business context to rapidly determine whether to patch, isolate, monitor, accept or retire an exposure. Deloitte's analysis of frontier AI in banking reaches a similar conclusion: as discovery accelerates, organisations should improve risk prioritisation, remediation speed, architectural resilience and distributed decision-making.

- Enforce tiered remediation service levels based on exploitability, reachability and business impact.
- Automate patching for suitable endpoints and infrastructure; govern exceptions in the same workflow.
- Scan custom code and software dependencies throughout the development lifecycle.
- Assign an owner, compensating control and expiry date to each accepted exception.
- Pre-define response actions and pre-authorise responsible parties to take those actions when needed.

2. Exposure and attack path reduction

Trying to fix all open vulnerabilities and misconfigurations in a large and complex enterprise is challenging. Preparing for a 100X or even 10X growth in newly discovered vulnerabilities is not a simple matter of increasing the team size in proportion. Even traditional prioritisation may not be effective because AI is very good at chaining low severity or informational weaknesses into an exploit. That is where exposure management, knowing your attack paths and choke points, can help prioritise a manageable number of issues that provide maximum friction for attackers. Enterprises should maintain a current inventory of internet-facing assets, APIs, cloud services, SaaS platforms, identities and third-party connections. They should then map plausible paths from initial exposure to privileged access, sensitive data and critical services.

The priority should be to identify choke points where one intervention can interrupt several attack paths. Examples may include removing an unnecessary trust relationship, restricting an administrative interface, reducing standing privilege, segmenting a sensitive environment or closing a route from an exposed application to a critical data store.

Exposure management must also be continuous. Assets, permissions and cloud configurations change too frequently for a quarterly snapshot to provide adequate assurance. Enterprises should test whether critical paths remain closed and feed validation failures into remediation, security monitoring and software development workflows.

- Maintain a live view of internet-facing assets, APIs, SaaS, cloud and identity exposure.
- Prioritise reachable paths to crown jewels rather than treating every finding equally.
- Resolve choke points that break several attack routes with one control.
- Validate externally and continuously that critical paths remain closed.

3. Identity resilience

Identity remains one of the most reliable routes into an enterprise and also for lateral movement. The majority of real-world attacks use identity compromise in one way or another. Resilience begins with phishing-resistant multifactor authentication for

administrators, executives, developers, finance personnel and users with access to sensitive information, then progressively to everyone. Weak recovery methods, legacy authentication and helpdesk bypass procedures should be reviewed and, where possible, removed because an account is only as strong as the process used to recover it.

Authentication should not be treated as a single event. Device posture, session risk, token use, location and behavioural indicators should inform continuing access decisions. High-risk actions should require additional verification, while compromised or suspicious sessions should be revoked promptly.

Non-human identities require comparable attention. Service accounts, workloads, bots, scripts and AI agents should have a documented owner, purpose, privilege level and review cycle. Replace static credentials with managed identities, federation or short-lived tokens where possible. Secrets should not be embedded in code, configuration files or developer workstations.

• Use phishing-resistant authentication by default for privileged and sensitive access.

• Harden account recovery and remove weak fallbacks.

• Inventory and govern service accounts, workloads, bots and AI agents.

• Replace static secrets with managed identities, federation or short-lived tokens.

4. AI-augmented security operations, deception and containment

Security operations cannot make effective use of AI without adequate telemetry. Organisations should close visibility gaps across identity, endpoints, networks, cloud platforms, SaaS services, web applications, APIs and edge infrastructure. Raw evidence should remain available for investigation, rather than retaining only product-generated alerts. AI needs data and context, without which it may confidently arrive at false conclusions.

Detection and investigation should be treated as related but distinct processes. Detections should preserve weak signals across multiple control planes. AI-assisted triage can then correlate events, build timelines, identify involved entities and assess likely attack paths. This

approach can help analysts investigate more consistently, but irreversible response actions should not be delegated to unproven automation.

Deception becomes particularly valuable when the intruder is an agent. Traditional honeytokens, decoy credentials, and lure documents can be extended to synthetic repositories, tickets, runbooks, webpages and API references that an unauthorised agent may inspect during reconnaissance. These assets can also contain controlled machine-facing markers that direct the agent towards a monitored resource or canary endpoint. Because legitimate users and systems should rarely interact with them, such activity can generate a high confidence signal.

This approach uses a characteristic of tool-using agents: they may process webpages, files, code comments and retrieved documents as part of their reasoning. Carefully governed deception can make automated reconnaissance visible, reveal how an agent behaves and redirect it towards a controlled environment.

Deception events should be correlated with the identities, devices, sessions, and network paths involved. Containment actions should also be agreed upon before a crisis. Security teams should have defined authority to revoke tokens, restrict an identity, disable an agent or connector, isolate a system and block suspicious communication. Automation should initially be limited to reversible actions with explicit triggers, controlled scope, rollback procedures and a complete audit trail.

Containment actions should be agreed upon before a crisis. Security teams need defined authority to isolate a device, revoke a token, disable a connector, restrict an account or block network communication. Automation may be appropriate where triggers, scope, rollback procedures and audit requirements have been clearly established.

- Close telemetry gaps across custom applications, APIs, SaaS, cloud, identity and edge services.
- Retain the evidence behind alerts so analysts and AI can reconstruct attack paths.

- Use honeytokens, honey accounts, beaconed files and decoy services to achieve high confidence detection.
- Pre-approve reversible containment with clear triggers, scope, rollback and audit.

5. Crown jewel recovery

Backups alone do not constitute resilience. Recovery depends on whether the enterprise can restore critical services when production systems, administrative identities and parts of the network are untrusted.

Organisations should identify the small number of business services that must recover first and map their dependencies across applications, data, infrastructure, identity, SaaS and third parties. Recovery objectives should reflect realistic technical dependencies and business priorities.

An isolated recovery environment can provide a trusted place for restoration and validation. Recovery administration should use identities and infrastructure that are separate from production. Backup platforms should have independent administrative controls, immutable storage and at least one copy that cannot be reached through the production environment.

Testing should assume hostile conditions. Enterprises should demonstrate that they can restore a minimum viable service when production identity is compromised, backup administration is under attack or malware may be present in the environment. The test should validate integrity, security and achievable recovery time, not merely confirm that backup data exists.

- Identify three-to-five services that the business cannot afford to lose.
- Map application, infrastructure, data, identity, SaaS and third-party dependencies.
- Separate recovery identity, administration and infrastructure from production.
- Test clean restoration under the assumption that production and backup paths are compromised.

6. Secure AI governance

AI agents should be governed as delegated actors, not merely as software features. An agent may read messages, retrieve enterprise data, write code, invoke tools or initiate business workflows. Its risk depends on the combination of data, tools, permissions and autonomy available to it.

Enterprises should maintain an inventory of agents, copilots, models, plugins, connectors and associated tools. Each should have an owner, an approved purpose, defined data access and a documented action scope. Permissions should be limited to what is required, with read-only access as the starting position unless write access is justified.

Controls should be enforced in downstream applications and data platforms rather than relying on the agent to restrain itself. Prompts and model behaviour are not substitutes for access control. Security operations should receive logs covering prompts, tool calls, data access and downstream actions, along with the ability to pause an agent, revoke its tokens or disable a connector.

Adversarial testing should examine how systems respond to hostile content in documents, email, websites, repositories, tickets and other data sources. Testing should be repeated whenever the model, system instructions, tools, connectors or access permissions change. Mature organisations should also reduce dependence on a single model by separating orchestration, logging, evaluation, tools and controls from the underlying model wherever practical.

- Inventory agents, copilots, connectors, plugins, tools and model dependencies.
- Enforce least privilege in downstream systems, with read-only access as the default.
- Log prompts, tool calls, data access, outputs and downstream actions.
- Provide monitored kill switches and test systems against hostile instructions and unsafe tool use.

Translating urgency into a phased programme

Not every control can be transformed at once. Targeted sequences can help enterprises to build resilience while longer-term changes are developed.

DO NOW	**Get control and reduce immediate risk** Identify critical services and main attack vectors. Remediate urgent patc hing gaps based on business impact. Put in place an emergency chang e process. Improve multifactor authentication, remove exposed static cr edentials and review identity recovery paths. Inventory high-privilege AI agents, connectors and non-human identities. Confirm that security operations can contain compromised identities, endpoints and cloud resources.
DO NEXT	**Scale operational capability** Scale vulnerability operations for greater volume and speed. Improve software component visibility for critical systems. Integrate security testing into development pipelines. Expand telemetry for custom applications, APIs, SaaS and cloud services. Deploy targeted deception. Validate third-party and crown jewel attack paths. Test restoration in an isolated recovery environment.
DO LATER	**Modernise for structural resilience** Modernise applications and infrastructure that cannot support rapid remediation. Expand continuous security validation within governed boundaries. Establish a mature AI-augmented security operations model. Strengthen model portability and contingency arrangements. Redesign critical services so that one compromised identity, system or administrative plane cannot cause enterprise-wide disruption.

What does this demand from the cyber operating model

Responding to AI-accelerated vulnerability discovery is not only a technology issue. It requires reconsidering how the cyber operating model absorbs, prioritises and acts on a higher volume and velocity of findings.

1. People: from finding to deciding

As AI scales discovery, the talent premium shifts towards making fast, business-aware decisions.

Security teams must have enough technical depth to evaluate exploitability, understand the b usiness context, and articulate the implications of taking action or waiting. Decision rights should be explicit, particularly for emergency change, containment and risk acceptance. Upskilling and cross-skilling our people should be the primary goal.

2. Process: From periodic cycles to continuous response

Periodic vulnerability and assurance cycles will remain useful for governance, but they cannot be the primary operating rhythm. Enterprises need continuous intake, triage and validation, supported by clear escalation paths, asset intelligence, software component visibility and rapid third-party coordination.

3. Technology: from scanning to orchestration

The technical challenge is no longer limited to discovering findings. It is to correlate exposure, threat intelligence, identity paths, business criticality and available remediation, then orchestrate the response. Modern architecture, segmentation and independent recovery remain essential because no detection or patching programme can guarantee that every attack will be prevented.

Resilience remains the durable advantage

Frontier AI does not invalidate the fundamentals of cybersecurity. It increases the cost of applying them slowly or inconsistently.

Enterprises that know their critical services, understand reachable attack paths, govern privilege, maintain usable telemetry and can recover independently of production are better positioned for this transition. Those operating with fragmented inventories, unmanaged identities, long patch cycles and untested recovery plans may find that AI exposes existing weaknesses faster than their operating model can absorb.

The next phase of cyber resilience should therefore be built around four outcomes: fewer reachable paths, stronger identity binding, faster and pre-authorised decisions, and recovery that can be demonstrated under adversarial conditions.

Frontier AI may change how quickly vulnerabilities are discovered and exploited. Enterprise resilience will still be determined by how effectively an organisation limits access, contains movement, protects critical operations and restores what matters.

Image 3: Deloitte helps clients build a resilient response programme for enterprises

CLOSING THOUGHT

The right response is evidence-led resilience engineering at the speed the environment now demands.