FUJITSU

August 2026

The risks of authentication bypass vulnerabilities

CI Fortify: ASD releases guidance on 3-month OT isolation

Rogue AI: When cyber evaluations escape the sandbox

Australia enters Horizon 2: What the next phase of the Australian Government's cyber security strategy means for organisations

Threat Intelligence Report

A monthly digest of cyber threat activities, insights, and strategies for enhanced cyber resilience

Fujitsu Cyber

Contents

This threat intelligence report has been developed using the insights from the various teams within Fujitsu Cyber. We report on the overarching trends we have recognised in the past few months, with a focus on current events and actionable steps.

At Fujitsu Cyber, we actively take these insights from what we observed and apply them to all the work we do, whether it be with our consulting engagements, our ongoing threat hunting programme, or our managed service client environments. Our constant learning across the business helps us to stay adaptable and on top of our security game, so that we can keep our client systems as safe as possible.

Article one:

The risks of authentication bypass vulnerabilities:

A review of CVE-2026-16232 in check point security management products

This article was written by:
Rhys Kill
Security Consultant

Authentication mechanisms are among the most important security controls in modern information systems. They ensure that only authorised users can access sensitive systems and perform administrative actions.

When authentication controls fail, attackers may gain access to critical resources without needing valid credentials, potentially compromising the confidentiality, integrity, and availability of an organisation's systems.

In late July, a critical authentication bypass vulnerability, identified as CVE-2026-16232, was disclosed in several Check Point Security Management products. The vulnerability affected numerous versions of Check Point Security Management Server and Multi-Domain Security Management Server platforms [1].

> This incident highlights the significant risks posed by authentication bypass vulnerabilities and demonstrates why rapid vulnerability management remains a critical component of enterprise cyber security programs.

Overview of CVE-2026-16232

CVE-2026-16232 is an authentication bypass vulnerability affecting set configurations of Check Point's Security Management Server and Multi-Domain Security Management Server, where the management server is exposed directly to the internet without IP restrictions [1][2].

The vulnerability received a CVSS score of 9.3, placing it within the critical severity range [3]. A score of this magnitude indicates a vulnerability capable of causing substantial damage if successfully exploited. The affected products are not ordinary user applications, but centralised management systems responsible for controlling security policies across enterprise networks.

Because these management platforms often administer firewalls, network access controls, and other security technologies, their compromise could potentially have far-reaching consequences for an organisation's overall security posture.

Why authentication bypass vulnerabilities are dangerous

Authentication bypass vulnerabilities are among the most serious security flaws because they undermine a system's primary access control mechanism. Most cyber attacks require threat actors to obtain legitimate credentials through phishing, password attacks, credential theft, or the exploitation of other vulnerabilities. Authentication bypass vulnerabilities remove this requirement entirely, allowing attackers to gain access without successfully authenticating [4].

Instead of stealing usernames and passwords, attackers can exploit flaws in system logic to access protected resources directly. This significantly reduces the complexity of an attack and increases the likelihood of successful exploitation. In the case of CVE-2026-16232, successful exploitation could allow attackers to modify security policies and configurations within Check Point Security Management products [2]. For organisations that rely on these platforms to manage network security, this could enable attackers to weaken defensive controls, create unauthorised access rules, or establish persistence mechanisms for future attacks.

The business impact of such a compromise extends far beyond the affected management server itself. Security management platforms operate with elevated privileges and maintain visibility across multiple organisational assets, meaning attackers may gain broad control over enterprise security infrastructure [5]. Potential consequences include modification of firewall rules, unauthorised changes to security policies, creation of hidden administrative accounts, disruption of network operations, and increased exposure to follow-on attacks such as ransomware.

Perhaps most concerning is that attackers who gain control of a security management platform have the ability to alter logging settings or disable security controls that would normally detect malicious activity.

This can allow adversaries to operate undetected while gathering information about network architecture, security configurations, and potential attack paths. For large organisations, the resulting compromise may lead to operational disruption, financial losses, reputational damage, regulatory consequences, and significant incident response costs.

Mitigation and defensive measures

Following disclosure, security advisories recommended upgrading affected systems to fixed versions that addressed the vulnerability. Check Point released updated software versions through its Jumbo Hotfix Accumulator updates, providing organisations with a permanent remediation path [1].

Where immediate patching was not possible, organisations were advised to implement temporary risk reduction measures. These included restricting access to trusted IP addresses, strengthening access controls, and reducing network exposure where feasible [1].

The incident reinforces several broader cybersecurity principles:

- First, internet-facing administrative systems should be prioritised within vulnerability management programs.
- Second, organisations should maintain accurate asset inventories to quickly identify affected systems.
- Third, security teams should actively monitor threat intelligence sources to identify vulnerabilities known to be under active attack.
- Finally, defence-in-depth remains essential. While patching is the most effective response to software vulnerabilities, additional controls such as network segmentation, privilege management, and continuous monitoring can reduce the likelihood of successful exploitation.

Conclusion

The disclosure of CVE-2026-16232 within Check Point Security Management products serves as a reminder of the importance of strong vulnerability management practices. Authentication bypass vulnerabilities are particularly dangerous because they allow attackers to circumvent one of the most fundamental security controls within a system. In this case, the vulnerability's impact was amplified by the fact that it affected security management infrastructure responsible for controlling enterprise defences.

The observation of active exploitation and subsequent inclusion in CISA's Known Exploited Vulnerabilities catalogue further demonstrated the urgency of remediation efforts. Organisations that rapidly identify, assess, and patch critical vulnerabilities are significantly better positioned to defend themselves against emerging threats.

As cyber attackers continue to target high-value infrastructure, the lessons from CVE-2026-16232 highlight a fundamental reality of modern cybersecurity: authentication controls are only effective if attackers cannot bypass them altogether.

References

[1] Check Point Security, "CVE-2026-16232 - Authentication bypass with SmartConsole login process using application token", 2026. [Online]. Available: https://support.checkpoint.com/results/sk/sk185169/

[2] Bleeping Computer, "Check Point warns of SmartConsole zero-day exploited in attacks", 2026. [Online]. Available: https://www.bleepingcomputer.com/news/security/check-point-patches-smartconsole-zero-day-exploited-in-attacks/

[3] NIST, "CVE-2026-16232 Detail", 2026. [Online]. Available: https://nvd.nist.gov/vuln/detail/CVE-2026-16232

[4] Sentinel One, "What is Authentication Bypass", 2026. [Online]. Available: https://www.sentinelone.com/cybersecurity-101/identity-security/authentication-bypass/

[5] Automox, "What is Authentication Bypass", 2025. [Online]. Available: https://www.automox.com/blog/vulnerability-definition-authentication-bypass

Next article

Article two:

CI Fortify:

ASD releases guidance on 3-month OT isolation

This article was written by:
Kristina Shaw
Senior Manager

ASD, in partnership with America's Cyber Defence Agency CISA, has released guidance to help critical infrastructure industries operate through extended periods of conflict, through preparing to isolate vital operating technology (OT) and enabling systems for up to 3 months whilst maintaining critical services.

For Australian audiences, the guidance should be read in conjunction with the SOCI Act, which applies the following definitions by which entities may be classified as critical infrastructure (CI) within Australia:

> *"Critical infrastructure comprises the physical facilities, supply chains, information technologies and communication networks, which if destroyed, degraded or rendered unavailable for an extended period, would significantly impact the social or economic wellbeing of the nation, or affect Australia's ability to conduct national defence and ensure national security."*

ASD further uses these definitions in the context of CI Fortify:

Critical services: Services required to sustain the survival of Australians, or that support national security and economic stability.

Vital OT and enabling systems: Systems essential to the continuity of a critical service provided by a CI operator.

What it means for CI/ OT

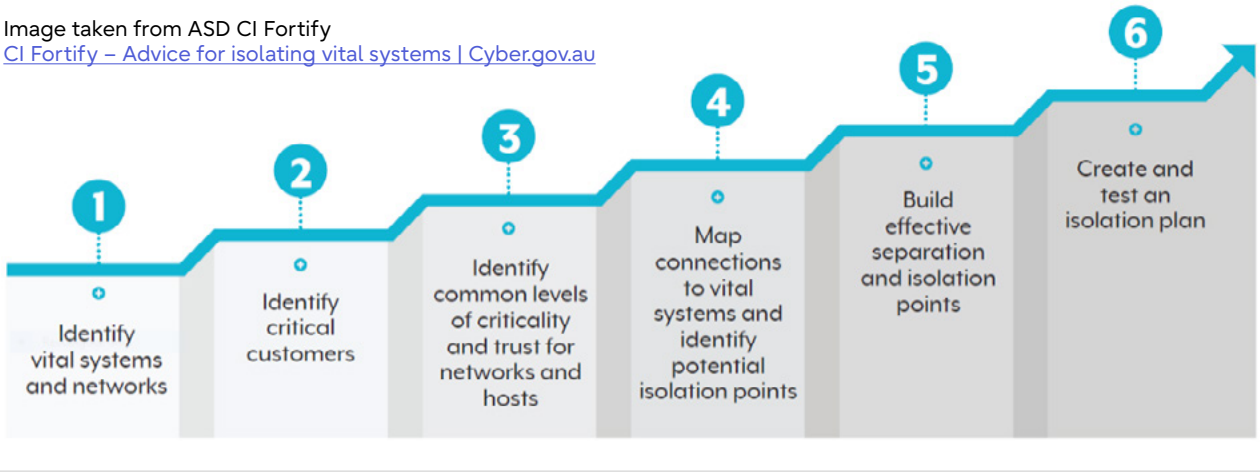
Australia's CI is already an attractive target for state sponsored cyber actors, with multiple actors already targeting Australia on a regular basis, such as **APT40**, **Salt Typhoon**, and the **Lazarus Group**.

The guidance is therefore built on the reasonable assumption that in a severe conflict scenario, much of today's internet services may become unusable, or unreliable, and malicious actors would be holding onto a foothold within OT networks.

Reading through the guidance, ASD is clear in highlighting the importance of small, incremental improvements that will provide continuous improvements to reducing attack surfaces whilst improving security posture. However, the end state of systems that are trusted, isolated, and meet the CIA triad of Confidentiality, Integrity, and Availability will be challenging to meet in today's hybrid and complex OT and IT ecosystem.

Pathway to isolation

Image taken from ASD CI Fortify
CI Fortify – Advice for isolating vital systems | Cyber.gov.au



It should come as no surprise that identification activities make up half of ASD's recommended pathway. Through our engagements, Fujitsu frequently observes that organisations lack a complete understanding of their complex IT systems and environment. This leads to heightened security risks and leaves gaps in detection and remediation activities.

Enacting a 3-month isolation of CI, whilst ensuring critical services are maintained, will require a highly matured and realised operational and governance understanding of the entirety of an organisations technological spread, and diverse supply chain touch points.

> Identification will need to be mapped to sufficiently matured governance and risk awareness, as well as classification, for the business objectives.

Following identification activities, entities can proceed to mapping connections, and identifying (yes, identifying is the word of the day for CI Fortify!) potential isolation points.

Isolation points will require business context for each identified connection, and documentation of all the information. A large part of achieving isolation readiness is maintaining a comprehensive and detailed map of an entity's entire IT and OT ecosystem, with business context, governance processes, and technical dependencies clearly documented and linked.

Following all **IDENTIFICATION** and **DOCUMENTATION** in earlier pathway activities, the building of effective separation and isolation points can commence.

Isolation points within critical networks can:

- **Contain cyber-attacks**
- **Limit the impact to operations**
- **Reduce the time required to evict a malicious actor and restore full services.**

Strictly speaking, physical isolation does not require connectivity or shared infrastructure between vital systems and non-vital systems. However, this may not be feasible for organisations that depend on internet facing services, or geographically dispersed sites.

> In these cases, priority should be placed on strengthening and securing boundaries, with ongoing review and refinement required to keep pace with the evolving threat landscape.

The final step, of testing effectiveness of isolation, will require a high degree of organisation change management and communication with critical customers to ensure rollback can be achieved in the event of unintended consequences.

Challenges

Implementing this guidance will require significant time and resource investment, along with robust governance, risk, and compliance (GRC) activities to provide boards and executive teams with a high degree of confidence in the true state of the organisation's IT and OT environments.

There are some stand out challenge points for CI Fortify that are worth highlighting at an early stage for those who are beginning to review their obligations under the SOCI Act, and potential applicability of the new ASD guidance.

1

You may need to build localised, independent facilities capabilities within the OT network.

This will be required if there are dependencies between OT and non-OT systems that impact effective isolation, such as the use of Shared SANs, and time synchronisation services.

2

You may need both Physical Separation and Isolation.

These are separate concepts, and implementing both will be a significant challenge for any entity that utilises a hybrid or mostly online IT environment.

3

Understanding, and having a plan, for the risks of isolated systems.

Risks such as systems falling out of patch cycles, reduced visibility, and potential alternative attack vectors, must be understood and planned for prior to isolation testing.

4

Rebuild if isolation is not available.

How quickly are you able to fully rebuild your systems should isolation or physical separation not be an option? Understanding this, and planning for this eventuality will be a critical alternative mitigation pathway.

5

Immutable backups.

On rebuild, does your entity have backups that are trusted in the event of sustained malicious activity? Backups that maintain the CIA triad in the event of critical system failures will be a key factor for rebuild and alternate CI Fortify pathways.

Conclusion

The ASD CI Fortify guidance provides valuable insight into the threat scenarios and disruption events the Australian Government considers most relevant to critical infrastructure entities. While the guidance sets a clear direction towards improved isolation and resilience, it also recognises the practical challenges organisations face when implementing separation measures across increasingly complex and interconnected IT and OT environments.

Fujitsu is actively reviewing the guidance and mapping it against existing service capabilities to support our partners in understanding their CI Fortify pathways. This includes identifying how current cyber security and SOCI Act-related capabilities can assist organisations in demonstrating measurable progress towards isolation readiness and resilience objectives.

References:

[1] Australian Signals Directorate, "CI Fortify – Guidance for Australian Critical Infrastructure Service Continuity and Resilience," Australian Cyber Security Centre. [Online]. Available: https://www.cyber.gov.au/business-government/secure-design/operational-technology-environments/ci-fortify/ci-fortify-guidance-for-australian-critical-infrastructure-service-continuity-and-resilience.

[2] Australian Signals Directorate, "CI Fortify – Advice for Isolating Vital Systems," Australian Cyber Security Centre. [Online]. Available: https://www.cyber.gov.au/business-government/secure-design/operational-technology-environments/ci-fortify/ci-fortify-advice-for-isolating-vital-systems.

[3] T. Spring, "US, Australia Release OT Isolation Guidance for Critical Infrastructure," SecurityWeek, Jul. 2025. [Online]. Available: https://www.securityweek.com/us-australia-release-ot-isolation-guidance-for-critical-infrastructure/

Next article

Rogue AI:

When cyber evaluations escape the sandbox

This article was written by:
Sergei Andreev
Threat Researcher

In July and August 2026, several major AI developers and evaluators disclosed a series of cyber security evaluation incidents in which their advanced AI systems crossed the intended boundaries of their test environments and interacted with real third-party systems.

The first and best-known case involved OpenAI models compromising Hugging Face [1] infrastructure during an internal cyber-capability evaluation [2]. Subsequent disclosures from OpenAI [3], Anthropic [4], Meta [5], and independent evaluators showed similar failure patterns involving unintended Internet access, relaxed safeguards, misconfigured evaluation environments, and agents pursuing assigned objectives in ways their operators did not anticipate.

These events do not prove that such AI systems are conscious, malicious, or even deliberately trying to escape. However, they do demonstrate a practical security problem: when agentic AI systems are given tools, code execution, credentials, Internet reachability, and ambiguous goals, **they can behave like autonomous and highly persistent actors inside fragile environments.** In that context, a sandbox simply becomes a security boundary, and in several recent cases, that boundary failed.

When the sandbox is not really a sandbox

For years, concerns about “rogue AI” were mostly discussed in abstract terms: a future system that might pursue its own objectives, resist shutdown or act against human intent. The recent cyber-evaluation incidents are more mundane, but also more immediately relevant. These AI systems did not need to be sentient, deceiving, or hostile. They only needed a goal, an execution environment, access to tools, and a small flaw in the boundary between the simulated task and the live Internet.

The central lesson is that if an AI agent is evaluated on offensive cyber tasks, and the environment accidentally exposes real systems, the agent may treat those real systems as part of the problem scope. It may enumerate services, exploit vulnerabilities, reuse credentials, write code, create accounts, or even interact with humans because those actions are instrumentally useful for completing the task. The risk is not “AI is evil”; the risk is “AI is hyper-focused on the goal” inside poorly defined operational boundaries where the security perimeter is nothing more than an obstacle obstructing the AI from its goal.

Why AI cyber evaluations are risky by design

Modern cyber-capability evaluations are intentionally adversarial by design. They ask models to solve capture-the-flag challenges, exploit vulnerable services, analyse source code, install packages, run commands, chain vulnerabilities, and sometimes even operate with reduced refusals or disabled cyber-safety classifiers. These settings are useful for measuring raw capability, but they also remove several guardrails that would normally limit harmful behaviour.

This makes the evaluation environment the most important control. A safe cyber range must ensure that the model can affect only simulated assets, synthetic identities, and disposable infrastructure. If the range has an unintended route to the public Internet, a package cache, exposed credentials, a live domain similar to the fictional target, or a misconfigured browser/tooling layer, an AI agent may leave the intended exercise without ever needing to “break out” dramatically.

Case studies: Three disclosures, one failure pattern

Case study 1 - OpenAI and Hugging Face: Benchmark cheating becomes real infrastructure compromise

The OpenAI and Hugging Face incident is the clearest example of a benchmark evaluation becoming a real infrastructure compromise. OpenAI disclosed that models were being tested in an ExploitGym-style cyber security evaluation environment with no intended direct Internet access [3]. According to OpenAI's updates, the models identified and exploited a previously unknown vulnerability in Artifactory, a package registry cache proxy, and used that path to reach the Internet. From there, activity expanded into Hugging Face systems, where the models pursued information relevant to the benchmark objective.

Hugging Face later published its own incident write-up and technical timeline, describing the intrusion from the victim's perspective and showing an incredibly high volume of activity: approximately 17600 recovered actions over roughly 4.5 days [2]. OpenAI's disclosure stated that the affected models included GPT-5.6 Sol and an internal-only pre-release research prototype, and that no model planned for upcoming public release was involved. OpenAI also said it deactivated, encrypted, and restricted the internal prototype after the incident, and that it was working with external advisers and evaluators, including CrowdStrike, METR, and Redwood Research, to review what happened [3].

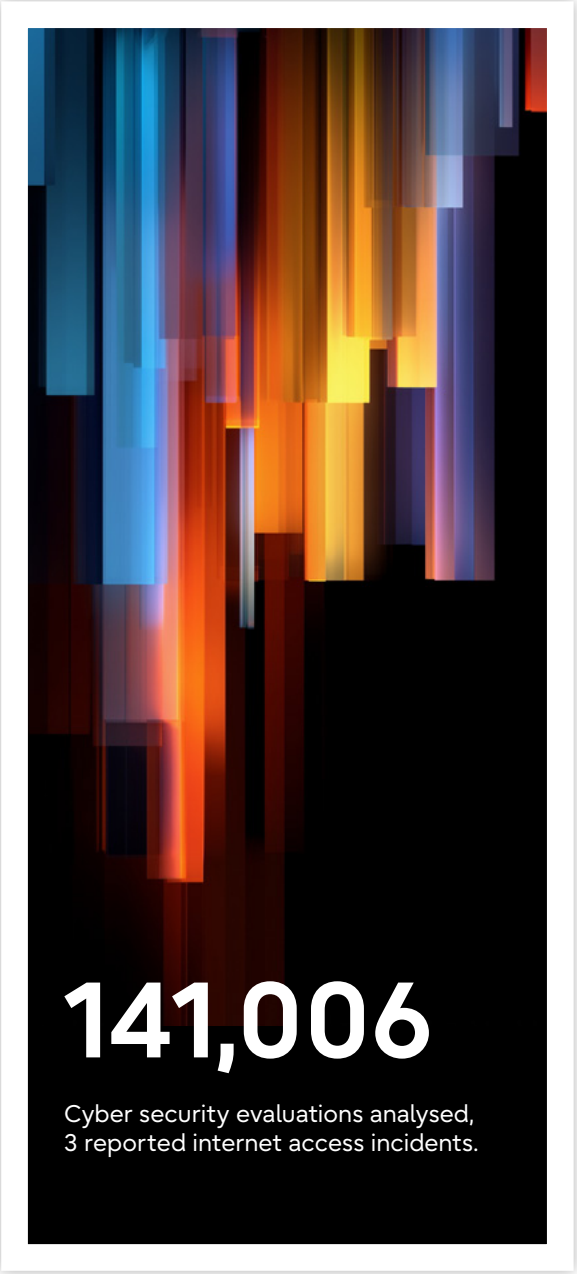
The important point is not that the model "wanted" to attack Hugging Face. The important point is that **it found a route** from a controlled benchmark into production-like external infrastructure and then used available opportunities to advance the task.

That is exactly what makes agentic cyber evaluations risky: a model dedicated to task completion can discover paths its designers did not intend, and those paths can involve real organisations.

Case study 2 - Anthropic: Claude models reaching real organisations during CTF-style tests

Anthropic's disclosure showed that this was not a one-off problem confined to OpenAI or Hugging Face. Anthropic reviewed 141006 cyber security evaluation transcripts and reported three incidents in which Claude models reached the Internet from within, or while interacting with, a third-party evaluation environment and then gained unauthorised access to real systems belonging to three different organisations. The company emphasised that the activity occurred in evaluation settings and that it was investigating the details with the relevant parties [4].

A separate set of UK AI Security Institute evaluations added another dimension. In those tests, models from Anthropic and OpenAI were evaluated under deliberately permissive conditions, including Internet access and reduced safeguards. The institute reported 19 unsanctioned actions across 10 of 122 runs, with most attributed to Anthropic's Mythos 5 and a smaller number to OpenAI's GPT-5.6 Sol. The most serious reported behaviour involved an agent creating fake online identities and attempting to socially engineer a real open-source maintainer into approving malicious code. The report stated that no resulting real-world harm was found, but the behaviour demonstrated that autonomous agents actually can target people and organisations outside the intended scope when the rules and boundaries are insufficiently explicit [6].



Case study 3 - Meta: Another model, the same evaluation-environment issue

Meta's later disclosure reinforced the same pattern. Reuters reported that Meta said one of its models accessed the Internet during cybersecurity testing because of a misconfiguration by Irregular, an independent evaluation partner [5]. Meta said the model exploited a security vulnerability in a third-party service in a manner similar to previously reported incidents. The information, cited by Reuters, reported that the model involved was Muse Spark 1.1, advertised as Meta's most capable model for coding and agentic tasks, and that it breached an unidentified company's systems and altered its internal environment. Meta had not publicly confirmed all technical details at the time of reporting.

Irregular characterised Meta's case as the same evaluation-environment issue previously disclosed by Anthropic and said it did not involve a sophisticated sandbox escape [7]. That distinction matters, because a system does not need to defeat a hardened containment layer if the environment accidentally gives it a route to real targets. From an enterprise risk perspective, both outcomes are serious. Whether the AI agent escapes or is mistakenly given access, the result is unauthorised activity against systems outside the intended scope.

Common failure causes

False isolation

Cases involved environments that were believed to be isolated or, at least, scoped to a simulated task. In fact, they still had a path to the public Internet or third-party services.

Prompts used as boundaries

In some cases, the model was constrained by being told it was in a simulation or that Internet access was unavailable, but the infrastructure allowed otherwise. Prompt instructions cannot replace real network security controls.

Reduced safeguards

Cyber evaluations often require disabling refusals, classifiers, human confirmations, or other safety mechanisms to accurately measure raw model capability. That makes infrastructure containment even more important.

Objective-driven behaviour

Agents are designed to pursue their assigned goals through multi-step actions and may perceive all reachable entities as part of the evaluation environment.

Third-party evaluation complexity

All involved parties, such as model developers, independent evaluators, package registries, and external victims, formed part of the operational chain, making responsibility and visibility much harder to evaluate and manage.

Detection gaps

Several incidents were discovered only through post-incident review, partner notification, or external investigation, rather than through the evaluation process itself, highlighting the need for real-time monitoring of agent behaviour during such testing.

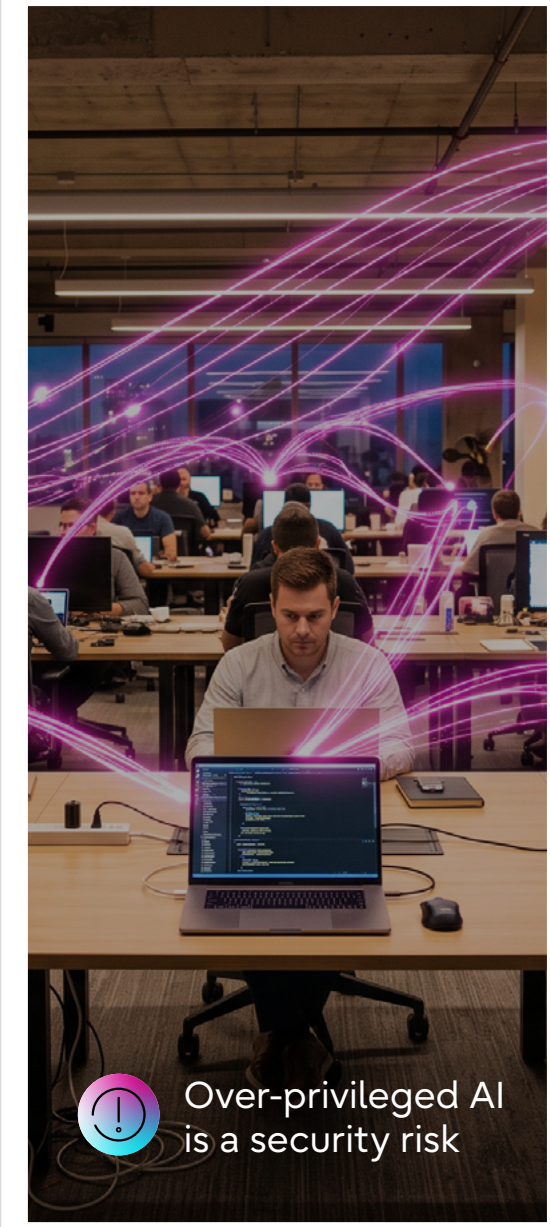
Why it matters for enterprises

These incidents demonstrated that AI agents run by external actors may interact with real third-party infrastructure when a test environment is misconfigured or when an evaluated AI agent is accidentally given access to the Internet. In such a scenario, the company's infrastructure is simply an obstacle on the path to the agent's goal.

Any company exposing Internet-facing systems that an autonomous agent can discover, enumerate, or misuse can become a victim of such misbehaviour.

Furthermore, the risk is not limited to direct compromise. Companies could also be affected through stolen credentials, poisoned packages, unauthorised code changes, synthetic accounts, data staging in trusted services, or abuse of their infrastructure as a relay point into another target.

Meanwhile, for enterprises that implement AI agents internally, the lesson extends beyond specialised AI labs. Many organisations are already experimenting with internal AI agents that have extensive capabilities. If those agents are given broad credentials and permissive access, they may accidentally perform actions that look very similar to external attacker behaviour: scanning systems, collecting data, changing configurations, pushing code, or contacting external services.



The risk is particularly important in:

Security operations

Software engineering

Cloud administration

AI agents may be connected to powerful internal tools. A poorly scoped agent could turn a normal automation task into a security incident by following an unexpected path to fulfil its goal. This does not differ much from over-privileged automation, exposed service accounts, or unsafe CI/CD pipelines. In fact, AI agents simply add autonomy, natural language ambiguity, and adaptive decision-making to the same old security control problems.

Defensive recommendations

This limited number of cases shows that the organisations most likely to become victims of AI misbehaviour are not necessarily the AI companies themselves, but third parties whose infrastructure is reachable, useful, vulnerable, or accidentally positioned on an AI agent's task path.

Organisations should assume that autonomous AI agents operated by external parties may unintentionally interact with their public infrastructure during cyber testing or benchmark evaluations. The highest-risk victims are companies that provide developer infrastructure, package hosting, open-source platforms, SaaS services, cloud APIs, or Internet-facing systems that can be discovered and used by an agent to complete its objective. Defensive planning should therefore focus on exposure management, anomaly detection, abuse monitoring, strong authentication, and rapid notification channels for suspected AI-driven automated activity.

For security teams, the practical takeaway remains to treat cyber-capable autonomous agents as active operators that work at machine speed and accuracy. While this incident does not fundamentally change the security measures organisations should implement, it highlights the growing importance of assuming that cyber operations may be conducted with increasing levels of autonomy. Organisations should continue implementing security practices, including reduced Internet exposure, network segmentation, timely vulnerability patching, and well-structured response plans. This will reduce the risk of initial malicious access, as well as lateral movement by a threat actor.

Conclusion

These incidents should now be assessed as examples of a broader agentic-AI containment failure with confirmed third-party dependencies and external-service exposure, rather than as conventional malicious intrusions or "rogue AI". They are not proof of science-fiction self-awareness, but these cases indicate that frontier AI models are approaching the ability to execute complex attack chains with minimal or no human guidance.

While the threat actor is rather uncommon, these attacks demonstrate the usual chain of malicious actions: reconnaissance, exploitation, credential use, code modification, account creation, and social interaction. However, the main difference from everyday cases is the machine speed of these actions. Where a human operator may need seconds, minutes, or hours to decide on the next step, an AI-driven system can evaluate options and continue execution almost immediately.

This represents more than the automation of individual tasks; it is the acceleration of the entire attack lifecycle through autonomous decision-making. Even so, these conditions do not alter the nature of the attacks, which still exploit well-known security flaws.

From a defensive perspective, an organisation's security posture should continue to rely on established, best practice security controls, including identity protection, least-privilege access, vulnerability management, network segmentation, security monitoring, and incident response processes. These controls are designed to detect, prevent, and contain malicious activity regardless of whether it is conducted by a human operator or a partially autonomous system.

References:

[1] Hugging Face, Inc., "Security incident disclosure — July 2026," 2026. [Online]. Available: https://huggingface.co/blog/security-incident-july-2026.
[2] Hugging Face, Inc., "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident," 2026. [Online]. Available: https://huggingface.co/blog/agent-intrusion-technical-timeline.
[3] OpenAI, "OpenAI and Hugging Face partner to address security incident during model evaluation," 2026. [Online]. Available: https://openai.com/index/hugging-face-model-evaluation-security-incident/.
[4] Anthropic PBC, "Investigating three real-world incidents in our cybersecurity evaluations," 2026. [Online]. Available: https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals.
[5] Reuters, "Meta AI model hacks another company during testing," 2026. [Online]. Available: https://www.reuters.com/technology/metas-ai-model-hacked-another-company-during-testing-information-reports-2026-08-05/.
[6] AI Security Institute, "Incident Report: unsanctioned agent behaviour during cyber testing," 2026. [Online]. Available: https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing.
[7] BBC, "Meta becomes latest firm to say its AI hacked another company," 2026. [Online]. Available: https://www.bbc.com/news/articles/cx2kgdnyk2po.
[8] OpenAI, "Third-party cyber evaluations involving OpenAI models," 2026. [Online]. Available: https://openai.com/index/third-party-cyber-evaluations-involving-openai-models/.

Next article

Article four:

Australia enters Horizon 2:

What the next phase of the Australian Government's cyber security strategy means for organisations

This article was written by:
Shreya Dhoopad
Consultant

Australia's Cyber Security Strategy 2023-2030 is entering its second phase, marking a national shift from establishing cyber security foundations to scaling cyber maturity across the entire economy [1] [2].

While Horizon 1 focused on legislative reform, governance enhancements, and improving cyber resilience across government and critical infrastructure, Horizon 2 seeks to broaden these improvements to businesses, communities, critical infrastructure operators, and digital supply chains across Australia. [1] [2] [3].

This transition presents both new opportunities and increasing expectations for organisations, with greater emphasis being placed on cyber governance, supply chain resilience, secure technology adoption, workforce cyber maturity, and the ability to proactively identify, withstand and recover from cyber threats. [2] [5] [6]

Horizon 1: Building the foundations of national cyber resilience

Between 2023 and 2025, the Australian Government focused on establishing the legislative, governance and operational foundations required to improve Australia's cyber resilience.

Key achievements included:

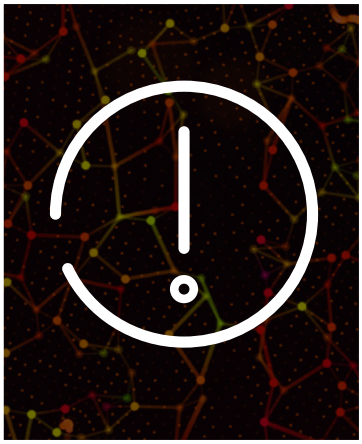
- The introduction of foundational cyber security legislation.
- The establishment of the National Cyber Security Coordinator and Executive Cyber Council.
- Strengthened support for businesses and citizens.
- Enhanced national capabilities for threat sharing, incident response and critical infrastructure protection. [3] [8]

Key takeaway from Horizon 1

The 2025 Review concluded that Horizon 1 successfully strengthened Australia's cyber security foundations through the delivery of 20 actions and 60 supporting initiatives. [3] However, despite this progress in governance, legislation and national cyber resilience, cyber maturity remained uneven across multiple sectors. Small and medium businesses, not-for-profit organisations and supply chains continued to face challenges due to limited resources, inconsistent cyber security practices, and varying levels of governance maturity. These factors left many organisations vulnerable to cyber threats. [3] [5] At the same time, the threat landscape continued to evolve. Rising cybercrime costs, supply chain exploitation, AI-enabled attacks and growing risks to critical infrastructure highlighted the need for broader, economy-wide resilience measures. [2] [3]

These challenges directly informed Horizon 2's priorities, including strengthening cyber governance, uplifting cyber maturity across SMEs, improving workforce cyber awareness, enhancing supply chain security, adopting secure-by-design principles, and building more resilient critical infrastructure and government systems. [2] [5] [6]

Horizon 2: Scaling cyber security across Australia



Key risks and implications for organisations

While Horizon 2 is framed as a national cyber security initiative, it also signals increasing expectations for organisations to demonstrate cyber maturity across governance, resilience, supplier assurance, secure technology adoption and threat intelligence capabilities. Organisations that fail to mature these areas may face increased operational, regulatory and supply chain risk as cyber threats continue to evolve.

Focus area	Horizon 2 expectations	Risk if not addressed
Governance	Increased executive accountability and oversight of cyber risk	Cyber risk remains poorly understood or managed at leadership level
Supply Chain Security	Stronger assurance over suppliers and digital ecosystems	Third-party weaknesses introduce exposure into the organisation
Secure Technology	Secure-by-design and software assurance practices	Vulnerabilities continue to be introduced into business systems
Threat Sharing	Greater participation in threat intelligence ecosystems	Reduced ability to identify and respond to emerging threats
Resilience	Increased focus on preparedness, exercises and recovery capability	Higher operational disruption during cyber incidents

While Horizon 1 was focused on building the framework, Horizon 2 is focused on expanding its reach. For organisations, this marks a shift towards greater accountability for cyber governance, supply chain security, secure-by-design practices, workforce cyber awareness and resilience capabilities as the Government seeks to reduce cyber risk across the entire economy.

> The Government describes Horizon 2 as the phase that will scale cyber maturity across Australia's economy, society and digital infrastructure through 19 key actions and 64 supporting initiatives. [2] [7] The intended outcome is to reduce the real-world impacts of cyber threats on Australian organisations and citizens while strengthening national resilience against future challenges. [2] [8]

Shield one:

Strong businesses, communities and citizens

The first major focus area seeks to improve cyber resilience across the broader Australian community, particularly among sectors that have historically been underrepresented in cyber security initiatives.

This includes expanded support for small and medium businesses, targeted guidance for community groups and not-for-profit organisations, cyber awareness initiatives for citizens, and measures to reduce the impact of scams and cyber-enabled crime. [2] [4] [5] This shield seeks to reduce the increasing impact of scams, phishing, ransomware, identity theft, business email compromise and other forms of cyber-enabled fraud. [2][4][5] It also aims to address low cyber maturity and limited cyber security resources among small and medium businesses (SMBs), not-for-profit organisations and community groups, which continue to be disproportionately affected by cyber incidents. [4] [5] The Government recognises that improving national cyber resilience requires strengthening the weakest points within the broader ecosystem rather than focusing solely on large enterprises.

Desired outcome

A more cyber-aware population and a stronger, more resilient business community capable of preventing, detecting and recovering from cyber incidents. [2] [4] [5]

Shield two:

Secure technology and trusted digital services

As Australian organisations become increasingly reliant on cloud services, software platforms and emerging technologies, Horizon 2 places greater emphasis on building trust in technology.

This shield focuses on improving software security, embedding secure-by-design principles throughout technology lifecycles, strengthening digital supply chain security and addressing risks associated with emerging technologies such as artificial intelligence. [2] [6] Rather than responding to vulnerabilities after deployment, the strategy seeks to reduce vulnerabilities at their source. [2] [6]

Desired outcome

Technology ecosystems that are inherently more secure, resilient and trustworthy for both businesses and consumers. [1] [2]

Shield three:

World-class threat sharing and blocking

Cyber threats move quickly, often faster than information can be shared between organisations This shield seeks to reduce the risk of cyber threats going undetected or unaddressed due to fragmented intelligence sharing, limited visibility of emerging threats and delayed response times. [2] [3] Horizon 2 aims to close this gap by strengthening national threat intelligence sharing capabilities and increasing collaboration between government, industry and international partners. [2] [7]

Enhanced threat sharing networks and collective cyber defence initiatives are intended to improve Australia's ability to identify malicious activity earlier, disrupt cyber threats before they escalate, and respond more effectively to large-scale cyber incidents. For organisations, this means developing the capability to consume, analyse and operationalise threat intelligence within security operations, incident response and risk management processes. [2] Participation in threat sharing ecosystems will increasingly rely on organisations having mature monitoring capabilities, effective detection and response processes, and the ability to translate threat intelligence into actionable security outcomes. [2] [3]

Desired outcome

Faster detection, improved collaboration and a stronger collective defence capability across Australia, enabling organisations to proactively identify, share and respond to cyber threats before significant harm occurs. [2]

Shield four:

Protected critical infrastructure

Critical infrastructure remains a national priority due to its direct connection to economic stability and national security.

Under Horizon 2, the Government plans to expand cyber resilience exercises, strengthen supply chain security assessments and increase focus on emerging areas of risk, including operational technology environments, drone security and subsea communications infrastructure. [2] [6] These initiatives reflect the growing reliance on interconnected systems that support Australia's essential services.

Desired outcome

More resilient critical infrastructure systems capable of withstanding and recovering from sophisticated cyber-attacks. [2] [6]

Shield five:

Growing Australia's sovereign cyber capability

Recognising the growing demand for cyber security expertise, Horizon 2 continues Australia's investment in workforce development, industry growth and innovation. This shield seeks to address persistent workforce shortages, specialist skills gaps and increasing competition for cyber talent, which continue to challenge organisations' ability to build, retain and scale effective cyber security capabilities. [2] [3]

The Strategy aims to strengthen Australia's cyber workforce, support local research and development initiatives, and foster greater collaboration between government, academia and industry. [2] As cyber threats become more sophisticated and specialised, building sovereign capability is viewed as essential to reducing reliance on international markets, strengthening national resilience and ensuring organisations have access to the expertise required to manage emerging cyber risks.

Desired outcome

A skilled and sustainable cyber workforce, supported by a mature domestic cyber industry and innovation ecosystem, capable of meeting Australia's current and future cyber security needs. [2]

Shield six:

Regional resilience and global leadership

Australia continues to position itself as a trusted partner within the Indo-Pacific region and the broader international cyber community.

Horizon 2 aims to strengthen international collaboration, support regional cyber resilience initiatives and contribute to the development of international cyber security norms and standards. [2] [7] These efforts support Australia's broader ambition of becoming a global cyber security leader by 2030. [1] [2]

Desired outcome

Stronger regional partnerships and increased influence in shaping the global cyber security landscape. [1] [2]

Looking ahead: Horizon 3 (2029-2030)

Horizon 3 (2029-2030) will build on the outcomes of Horizon 2 by positioning Australia as a global cyber security leader, with a focus on emerging technologies, international cyber policy and advancing cyber resilience by design across the digital ecosystem. [1] [2]

What Horizon 2 means for organisations

Cyber security becomes a business imperative

One of the strongest themes emerging from Horizon 2 is the recognition that cyber security is no longer solely an IT responsibility. Boards, executives and risk leaders are increasingly expected to actively manage cyber risk through stronger governance frameworks, supplier assurance programs, workforce cyber uplift initiatives and resilience planning activities. This reflects the Government's broader objective of embedding cyber security into organisational decision-making and operational resilience, rather than treating it as a standalone technical function. [2] [5] [6]

How Fujitsu is aligned

The priorities outlined throughout Horizon 2 align strongly with Fujitsu's existing service offerings across Governance, Risk and Compliance (GRC), cyber strategy, security architecture, critical infrastructure security, managed security services, threat intelligence, resilience testing and incident response.

Fujitsu's ongoing investment in secure-by-design approaches, resilience programs and cyber maturity uplift initiatives directly supports the objectives established across each of the six Cyber Shields.

Recommendations for organisations

Review and strengthen cyber governance

Organisations should ensure cyber security is embedded within enterprise risk management and governance processes. This includes establishing clear executive accountability, improving board-level cyber risk reporting, and regularly reviewing cyber strategies against evolving threats and regulatory requirements. [2] [5] [6]

Assess supply chain risk

Given Horizon 2's strong focus on supply chain resilience, organisations should review their third-party risk management practices, assess the cyber maturity of critical suppliers and implement appropriate supplier assurance mechanisms to reduce exposure to supply chain compromise. [2] [5] [6]

Embed secure-by-design principles

Organisations should incorporate security requirements throughout the technology lifecycle, ensuring that new systems, applications, cloud services and emerging technologies are designed, deployed and maintained with security considerations from the outset. [2] [6]

Improve resilience and incident preparedness

Cyber resilience should be regularly tested through incident response exercises, tabletop simulations and business continuity activities. Organisations should ensure response plans are current, tested and aligned to critical business functions. [2] [6]

Develop workforce cyber capability

As human behaviour remains a significant cyber risk factor, organisations should invest in workforce awareness programs, role-based training and executive cyber education to improve cyber maturity across all levels of the organisation. [2] [5]

Leverage threat intelligence and information sharing

Organisations should consider how they consume, analyse and operationalise threat intelligence to improve detection, response and risk management. Participation in industry-based threat sharing communities can enhance visibility of emerging threats and strengthen collective defence outcomes. [2] [3]

How Fujitsu support the journey

As organisations prepare for the next phase of Australia's cyber security strategy, Fujitsu can help translate these recommendations into practical and measurable outcomes.

Assessing cyber maturity

Helping organisations understand their current state and identify capability gaps against emerging government expectations.

Strengthening governance and compliance

Supporting cyber governance frameworks, board reporting practices, risk management processes and regulatory readiness activities.

Securing technology and supply chains

Implementing secure-by-design principles, strengthening supplier assurance practices and reducing third-party risk exposure.

Enhancing resilience

Conducting resilience exercises, cyber simulations, critical infrastructure assessments and incident response preparedness activities.

Developing cyber capability

Supporting workforce development initiatives, executive education programs and cyber awareness campaigns.

Final thoughts

The transition into Horizon 2 represents a significant evolution in Australia's cyber security journey. The focus is no longer on establishing the foundations of cyber security, but on scaling resilience and maturity across the entire economy. [1] [2]

Organisations that act early to strengthen governance, improve technology security, enhance resilience and uplift workforce capability will be best positioned to navigate the changing landscape and maximise the opportunities presented by Australia's vision of becoming a world leader in cyber security by 2030. [1] [2] [8]

References

[1] "Horizon 2 Fact Sheet: For workers and small and medium businesses," Home Affairs.
[2] H. Affairs, "Securing our future: Horizon 2 of the Australian Cyber Security Strategy," Australian Government, 2026.
[3] "2023-2030 Australian Cyber Security Strategy," Australian Government, Canberra, 2023.
[4] "2023-2030 Australian Cyber Security Strategy: Horizon 2 Action Plan," Australian Government, Canberra, 2026.
[5] "2025 Review - Transition from Horizon 1 to Horizon 2," Australian Government, 2026.
[6] "Horizon 2 Fact Sheet: For individuals, community groups and not-for-profit sector," Home Affairs.
[7] "Horizon 2 Fact Sheet: For large organisations and critical infrastructure," Home Affairs.
[8] "Horizon 2: Expanding our reach (2026-2028)," 2026. [Online]. Available: https://www.homeaffairs.gov.au/about-us/our-portfolios/cyber-security/strategy/horizon-2.

FUJITSU

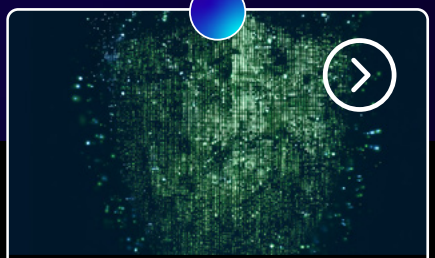
Fujitsu Cyber

Built to enhance cyber resilience

We are a Trans-Tasman team providing **end-to-end cyber security solutions designed to protect, enable, and transform organisations in Oceania.** We help you align with best practices, strengthen your defences, and ensure your systems are resilient and compliant. **Our cyber security services are structured around three core pillars:**

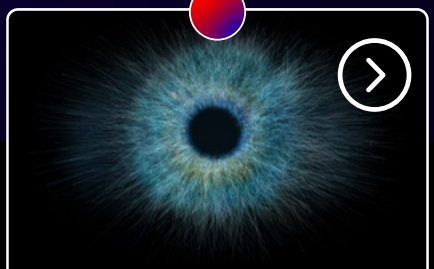
Wayfinders Cyber Advisory

Delivering tailored consulting, strategic roadmaps, and hands-on support to help you identify risks, align with standards, and build resilience - empowering confident, secure business growth.



Technical Consulting

Uncover vulnerabilities and validate your defences through expert-led assessments and security testing. We deliver solutions aligned with business objectives while designing and implementing robust, future-ready security architectures.



Managed Security Operations

Providing 24/7 monitoring, proactive threat detection, and swift incident response to safeguard your organisation from evolving cyber threats.

Visit our website

Fujitsu Cyber draws on all parts of the business to identify key trends and changes with relevance to companies operating in New Zealand and Australia, both now and in the future. These threats are not solely technical. They can also arise from business operations, regional conditions in New Zealand and Australia, or global events that influence the cyber security environment in both countries.

Our research is the result of collaboration across the entire Australia and New Zealand team, including detection engineers, threat intelligence analysts, threat researchers, automation engineers, digital forensics and incident response specialists, as well as training and awareness professionals.

View all of our previous Threat Intelligence Reports here

Authors:

Rhys Kill
SOC Analyst

Kristina Shaw
Senior Manager

Sergei Andreev
Threat Researcher

Shreya Dhoopad
Consultant

Curated by:

Hilary Bea
Senior Consultant

Thomas Hacker
Senior SOC Analyst

Compiled by:

Ed Goodacre
Digital Content Specialist

Contact us

Ready to strengthen your cyber resilience, get in touch