Australian Government
Australian Signals Directorate

cyber.gov.au

Opportunities for AI in cyber defence

Use of AI by cyber security teams

First published: May 2026

Last updated: September 2026





Australian Government

Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE

ACSC Australian Cyber Security Centre



Communications Security Establishment Canada

Canadian Centre for Cyber Security

Centre de la sécurité des télécommunications Canada

Centre canadien pour la cybersécurité

National Cyber Security Centre

NEW ZEALAND



National Cyber Security Centre

a part of GCHQ

Table of contents

Introduction **1**

Audience **2**

The cyber security landscape is evolving **3**

Spectrum of AI use in cyber security 3

Agentic AI 4

Malicious actors are leveraging AI 4

Integrating AI for cyber security **5**

Govern 5

Identify 6

Protect 7

Detect 8

Respond 9

Recover 10

Securely adopting AI **12**

Human oversight 12

System protection and sandboxing 13

Secure system integration 13

Governance 14

Supply chain 15

Testing and assurance of AI tools 15

AI Secure by Demand 16

Conclusion **17**

Further information **18**

ASD's resources 18

Other resources 18

Contact details **19**

Appendix A: Cyber security questions for AI vendors 20

Measurable security outcomes 20

Scope, capability and limitations 20

Human oversight and decision making 20

Transparency, explainability and auditability 21

Security of the AI system itself 21

Data protection and sovereignty 21

Integration and operational fit 22

Resilience, dependency and long-term risk 22

Appendix B: AI capability questions for cyber security vendors 23

Security outcomes and effectiveness 23

Scope, assumptions and constraints 23

Handling uncertainty and error 23

Transparency and auditability 24

Security of the AI system 24

Data use and protection 24

Integration and operational fit 24

Resilience and recovery 24

Introduction

Artificial intelligence (AI) is rapidly reshaping the cyber security landscape. As highly capable AI becomes more widely available, malicious actors are using it to deliver cyber threats at greater scale and speed. Organisations that do not re-evaluate and improve their defences will remain vulnerable to these AI-enabled cyber threats.

Cyber security has traditionally relied on specialised teams and reactive workflows to manage risk. These approaches remain important, but the scale and complexity of the modern cyber security landscape increasingly strain them. Heavy dependence on manual processes can make it difficult to prioritise risks, investigate potential threats and maintain consistent defensive coverage.

AI presents a significant opportunity for cyber defenders. When used safely, securely and responsibly AI can:

- strengthen prioritisation of cyber risks
- improve detection of threats and vulnerabilities
- support faster response and recovery
- reduce reliance on repetitive manual tasks.

This guidance was co-seal by the Canadian Centre for Cyber Security (CCCS), the New Zealand National Cyber Security Centre (NCSC-NZ) and the United Kingdom National Cyber Security Centre (NCSC-UK), and addresses how organisations can use AI to strengthen organisational cyber security while managing the risks of using AI. It outlines how the cyber security landscape is evolving and describes how organisations can use AI aligned with the *Information security manual* (ISM) cyber security functions of Govern, Identify, Protect, Detect, Respond and Recover. It also sets out principles for securely adopting AI, along with key questions for cyber defenders to ask AI vendors to support secure use.

Human oversight, governance and Secure by Design practices remain essential. AI can significantly enhance cyber security, but it is not a replacement for strong cyber security fundamentals. Poorly designed or poorly governed AI systems can introduce new attack paths. This can occur through excessive system access, reliance on untrusted inputs, or automated actions without adequate safeguards. Organisations that lack robust identity and access management, secure configuration, patch management, network segmentation, monitoring and incident response processes are unlikely to achieve better security outcomes simply by deploying AI.

Audience

This guidance supports cyber defenders, such as chief information security officers and senior security leaders who are responsible for cyber security strategy, operations and risk outcomes.

The publication assumes a foundational understanding of cyber security concepts and supports informed decision-making about safe, secure and responsible use of AI within cyber security.

The cyber security landscape is evolving

The cyber security landscape is rapidly evolving as malicious actors increasingly use AI to improve the speed, scale and sophistication of attacks. To adapt, cyber defenders can use AI to strengthen analysis, prioritisation and defensive decision making.

Spectrum of AI use in cyber security

Organisations are increasingly exploring how AI can be used to enhance cyber security outcomes. The focus should remain on improving security within the organisation, noting that simply identifying vulnerabilities does not, on its own, strengthen security. Without appropriate context, prioritisation and remediation, poorly implemented AI use could introduce additional risk rather than reduce it.

Organisations should draw on a spectrum of AI capabilities depending on their objectives, risk appetite and available technology. This spectrum ranges from advanced frontier AI models, through to general purpose large language models (LLM), to embedded AI features within existing security tools. For a list of questions to support vendor assessment, refer to Appendices A and B.

At the leading edge, frontier AI models can perform a wide variety of tasks and reflect the capabilities present in today's most advanced models. Compared to more common generative AI systems, frontier AI enables more complex reasoning, broader task coverage, and tighter integration with tools, data and operational workflows. Frontier AI use also carries cost and sustainability considerations, including token consumption, compute demands and ongoing operational expenditure. Organisations should consider these factors when selecting AI models, ensuring the level of model capability is commensurate with the requirements and value of the intended use case.

In a cyber security context, frontier AI should be integrated to augment existing tools and processes, rather than deploying it as a standalone solution. Within cyber security operations, organisations achieve safer and more effective AI use by deploying modern and fit-for-purpose security software that constrains, tests and governs AI capabilities to support specific ISM functions. The joint Five Eyes cyber security agencies statement urges organisation leaders to use AI to strengthen defence.

The approach an organisation takes will depend on what it is trying to achieve and what it has access to. A layered strategy allows organisations to combine these capabilities, using more advanced AI where appropriate while leveraging existing tools to enhance operational efficiency.

This ensures AI is used to strengthen existing cyber security practices rather than replace them, supporting safe, secure and responsible adoption across the organisation.

Not every organisation can adopt frontier AI at the same pace, especially where resources and skills are limited, creating gaps between cyber defenders. The World Economic Forum highlights that smaller and public-sector organisations face lower cyber resilience, with critical infrastructure showing particularly low confidence in responding to major incidents.

These challenges point to the need for stronger local readiness, trusted partnerships, and practical support.

Agentic AI

Agentic AI refers to AI systems that can independently plan, decide and take actions to achieve a goal, rather than simply responding to individual prompts. These systems use advanced AI models, such as LLMs, to understand their environment and reason about options. They are combined with access to tools, data, memory and workflows, enabling them to act and operate with a degree of autonomy.

Unlike traditional AI or chat systems, agentic AI actively works toward outcomes even when objectives are loosely defined. It can operate with limited ongoing human oversight, adapt its behaviour based on results, and in some cases create subtasks or subagents to complete complex work. While humans set goals, constraints and permissions, agentic AI systems execute goal directed actions over time rather than simply providing advice or outputs.

For more information refer to our co-authored publication by Cybersecurity and Infrastructure Security Agency (CISA), National Security Agency (NSA), Canadian Centre for Cyber Security (CCCS), National Cyber Security Centre New Zealand (NCSC-NZ), National Cyber Security Centre (NCSC-UK) on careful adoption of agentic AI services.

Malicious actors are leveraging AI

Malicious actors are increasingly leveraging AI to accelerate development of cyber attacks and deploy them at scale. They embed AI models directly into malicious workflows to automate reconnaissance, develop attack tooling, analyse compromised data and generate tailored malicious outputs. This accelerates vulnerability discovery, and shortens the time between vulnerability discovery and exploitation, leaving defenders with less warning and response time.

AI also lowers the technical barriers to entry. Less skilled malicious actors can use AI to perform activities that previously required specialist expertise, such as producing evasive malware, conducting large-scale data analysis or executing convincing social engineering campaigns.

As AI capabilities become more accessible, malicious actors will continue to increase the speed, scale and impact of malicious cyber operations. To stay ahead, organisations should strengthen defensive capability by reinforcing cyber security fundamentals. These include minimising attack surface, promptly patching systems, implementing layered defence-in-depth architectures, increasing automation and improving threat detection. Without corresponding advances in defence, AI-enabled threats are likely to erode the effectiveness of traditional security approaches.

For further information, refer to the United Kingdom's National Cyber Security Centre's (NCSC-UK) *Impact of AI on cyber threat from now to 2027* assessment.

Integrating AI for cyber security

The use of AI for cyber security should align with the ISM and its associated controls. The ISM groups cyber security principles into 6 functions: Govern, Identify, Protect, Detect, Respond and Recover. This section outlines how organisations can apply AI within each function to support cyber security, as illustrated in Figure 1.

Many AI use cases involve processing sensitive operational information. Organisations should apply AI while maintaining appropriate controls and protecting information, in line with their security and data handling obligations. This includes enforcing least privilege access, restricting data exposure to AI systems, validating outputs before use and maintaining auditability of AI-assisted actions.

Organisations should also undertake ongoing assessment of costs, benefits and risks, recognising that both AI capabilities and malicious actor use of AI will continue to evolve.

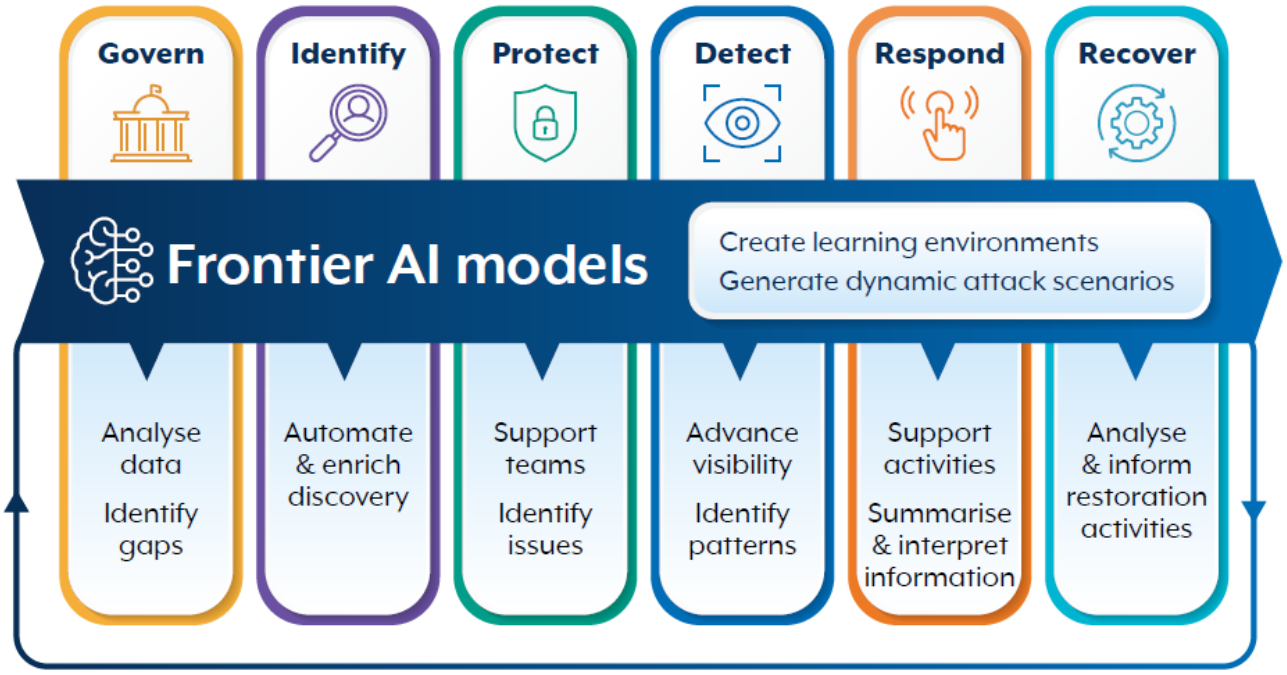


Figure 1: Applying AI across the six cyber security functions

Govern

Develop and maintain a strong and resilient cyber security culture

The *Govern* function focuses on how cyber security and cyber risk are directed, understood and managed across the organisation. It establishes clear accountability, decision making structure and oversight mechanisms that connect executive leadership with technical teams. As threats evolve, effective governance supports timely review and adaptation of policies, controls and assurance arrangements to maintain organisational resilience.

AI-specific threat intelligence sharing should also be strengthened, leveraging ASD's ACSC (including its Partnership Program), the National Cyber Threat Notification System (NCTNS), and CISA's AI Cybersecurity Collaboration Playbook to support timely, trusted collaboration across partners.

As organisations adopt AI in cyber security, governance should define clear expectations for its use, oversight, and human accountability. This includes applying bounded autonomy to keep AI within defined limits, especially in critical systems and security operations. It also requires recoverability so AI functions can be safely restricted or restored, ensuring operations continue if systems fail or are compromised.

AI may support organisations to:

- identify inconsistencies in risk evaluation across business units, systems or projects
- analyse supply chain risks, including software dependencies, vulnerability exposure and vendor security practices
- support the creation and use of inventories, such as software bill of materials (SBOM) and cryptographic bill of materials (CBOM)
- strengthen policy interpretation and compliance, such as using an internal AI assistant to provide context aware guidance based on organisational policies and regulatory requirements
- prioritise cyber security decisions based on risk assessment.

Identify

Identify assets and associated security risks

The *Identify* function focuses on understanding what assets exist within the environment and the security risks associated with them. It establishes visibility of systems, software, data and configurations to enable informed risk-based decisions.

AI may support organisations to:

- enhance asset discovery by using network telemetry to identify unmonitored or hidden assets across the environment
- prioritise patching decisions by using multiple factors including severity ratings, exploit availability, threat intelligence and operational impact of exploitation
- identify and assess vulnerability chaining by linking multiple lower severity vulnerabilities into attack paths
- identify insecure configurations and recommend remediation actions
- review log samples to identify actions an adversary could take without triggering alerts, and assist with refining detection logic

- analyse SBOMs and CBOMs to map all software components to identify outdated libraries, hidden risks, and supply chain exposures
- map AI training and fine-tuning data sources to assess provenance, retention and poisoning exposure.

Red team scenario: AI-driven attack path analysis

An organisation's red team uses a suitable AI model to analyse its enterprise environment, ingesting system architecture, identity relationships, and vulnerability data at scale. The model is configured with relatively broad permissions, allowing it to correlate findings across network, application and access control layers.

Over time, the AI identifies how multiple low- and medium-severity weaknesses can combine into meaningful attack paths that would be difficult and time-consuming for humans to detect manually. For example, it links a benign misconfiguration, excessive permissions and a minor software flaw into a sequence that could enable privilege escalation and lateral movement.

By continuously generating and refining these scenarios, the red team uses AI to uncover high-impact chains that are not obvious when issues are considered in isolation. This improves testing coverage and helps organisation prioritise remediation based on real-world attack feasibility rather than individual vulnerability severity.

For more information on SBOMs, refer to the United States' Cybersecurity and Infrastructure Security Agency's *A Shared Vision of Software Bill of Materials (SBOM) for Cybersecurity*.

Protect

Implement and maintain controls to manage security risks

The *Protect* function focuses on implementing and maintaining safeguards that reduce the likelihood and impact of cyber security incidents. This includes maintaining secure configurations, managing identities and access, reducing attack surface, and ensuring controls remain effective as systems and threats change.

AI may support organisations to:

- prioritise hardening actions based on exploitability and environmental context to reduce the attack surface
- enhance security architecture by evaluating complex environments, trust boundaries and data flows to simulate realistic attack paths and detect weaknesses
- analyse identities, roles and behaviours of users and agents to identify breaches of least privilege access, detect excessive or unintended permissions, and identify privilege creep or orphaned accounts

- analyse real-time traffic patterns to identify potential security threats, including anomalous or autonomous AI driven activity, and recommend improvements to network segmentation rules
- scan source code, infrastructure as code, and pipeline definitions to discover vulnerabilities and business logic flaws that pattern-based scanners miss
- zero-trust and modern defensible architectures should extend to AI agents as well as human users. This involves replacing static credentials with dynamic tokens and enforcing clear privilege boundaries on what agents can access and execute.

Code review scenario: AI identifying hidden vulnerabilities

An organisation deploys an AI tool to review source code that automatically analyses source code whenever a developer submits changes. During a routine update, a developer introduces a feature that processes user input and stores it in a database. While the code functions as intended, the AI identifies that the input is not properly validated before storage.

Although this issue appears minor and would likely pass traditional checks, since the syntax is correct and no obvious rule is violated, the AI recognises the broader security implication. It highlights that a malicious actor could exploit this gap to insert malicious data, potentially leading to compromise.

By flagging the issue early, the organisation is able to correct the validation logic before release. This reduces the risk of introducing exploitable vulnerabilities into production and strengthens overall security without slowing development time.

For further information on security evaluations of AI, refer to the Frontier Model Forum's *Technical Report: Managing Advanced Cyber Risks in Frontier AI Frameworks*, which outlines emerging industry views on how frontier AI models can support cyber risk management, including the identification of vulnerabilities and insecure defaults.

Detect

Detect and analyse cyber security events to identify cyber security incidents

The *Detect* function focusses on identification and analysis of cyber security events to enable the timely detection of potential incidents. It supports continuous monitoring across systems, networks and identities to surface anomalous activity and emerging threats.

AI may support organisations to:

- detect cyber security events and incidents, with analyst judgement to validate findings
- leverage frameworks such as MITRE ATLAS, OWASP Top 10 for LLM Applications, OWASP Top 10 for Agentic Application, NIST AI RMF Generative AI Profile and MAESTRO threat

model to inform AI-supported detection of malicious behaviours targeting AI-enabled systems

- analyse network telemetry (for example, flows, logs, Domain Name System, and application programming interface calls) to detect anomalous behaviour, such as unexpected communication between services or connections to known malicious endpoints
- detect AI misuse through AI-specific telemetry, such as model inputs (including prompt manipulation), decision traces, policy checks and confidence signals, while protecting the integrity of detection logs from tampering
- analyse behaviour and configuration context to distinguish legitimate activity from suspicious activity to minimise false positives and false negatives
- baseline high-risk activities such as identity, privileged access and remote access to assist in anomaly detection.

Security operations centre platform scenario: AI-driven threat detection and triage

An organisation has a security operations centre (SOC) platform that uses AI-assisted detection to triage alerts across identity, endpoint, network and cloud telemetry. To help analysts manage high alert volumes, the platform correlates data across the Domain Name System, application programming interface (API) transactions, and network flows. It also tracks patterns in privileged access and surfaces prioritised incidents for analyst review.

Overtime, the organisation builds confidence in the platform's ability to suppress false positives and highlight the alerts that matter most. The AI continuously refines its understanding of normal and abnormal behaviour, improving detection accuracy across the environment.

In this scenario, the SOC platform enables analysts to identify and respond to the most critical threats more quickly, reducing response times and alleviating operational overload while maintaining strong security visibility.

Respond

Respond to cyber security incidents

The *Respond* function focuses on effective, timely and coordinated action during cyber security incidents. It ensures organisations can contain and mitigate incidents while maintaining critical operations.

AI may support organisations to:

- assist analysts by correlating alerts, logs and forensic artefacts into a coherent explanation of the incident

- interpret alerts and observed system or user behaviours in the context of the organisation's specific systems, architecture and risk environment
- reduce reliance on manual searches during investigations
- sequence response actions across identity, endpoint and network controls
- draft incident updates that bridge responders and executives by aligning with defined communication expectations, with responders validating accuracy before distribution
- augment surge capacity by automating triage and preparing or initiating pre-approved and reversible response actions in parallel across concurrent incidents, with human approval required for high-impact containment, disruption, recovery or communication decisions, which is especially critical in AI-driven attack scenarios
- propose containment and remediation actions for uncertain cases, aligned with organisational incident response playbooks, for review by human incident responders.

Recover

Resume normal business operations following cyber security incidents

The *Recover* function focuses on restoring systems and services following a cyber security incident so normal business operations can safely resume. It prioritises and controls recovery activities based on verified system integrity and accepted residual risk.

AI may support organisations to:

- analyse rebuild and restoration pathways to support recovery planning, sequencing and assurance activities
- trigger automated recovery actions and playbooks to remediate impacted assets and return them to a secure, trusted state, with human approval for destructive or irreversible changes such as system restoration or data rollback
- validate system and service restorations against known baselines to confirm integrity before resuming operations
- rollback AI models to previous versions where compromise, poisoning or unintended model drift is suspected
- verify the integrity and expected behaviour of AI models and associated data through AI-specific validation checks before re-enabling AI-augmented or autonomous functions
- enable rapid and safe resumption of critical business services by identifying potential cascading failures across interconnected systems and reducing the risk of occurrence
- generate and evaluate recovery sequences for novel or complex incident types that fall outside predefined automation or security orchestration, automation and response playbooks

- identify weaknesses in recovery arrangements, dependencies or assumptions proactively, before they are encountered during a real incident.

Securely adopting AI

When adopting AI, organisations should prioritise strong security practices. While no mitigation strategy can provide complete protection, organisations should implement a strong cyber security baseline aligned with ASD's ISM to materially reduce cyber security risk.

As AI capabilities and the cyber threat environment evolves, organisations should regularly review and update their controls to ensure AI continues to be used safely, securely, responsibly and in accordance with organisational risk tolerance. Critical infrastructure (CI) should place particular focus on maintaining resilient, continuously assured controls given its heightened risk and impact.

Data readiness is a core prerequisite for AI adoption. AI-supported cyber defence will only be as reliable as the underlying data it draws on, including asset inventories, logging coverage, identity and privilege data, vulnerability records and configuration data. Organisations should assess and uplift the quality, coverage and currency of this data before and during AI adoption, recognising that gaps or inconsistencies will directly limit the accuracy and value of AI-generated insights.

Adopting AI in cyber defence depends on workforce capability as much as technology. Cyber defenders should understand how to use AI tools effectively, including how to validate outputs, recognise limitations and avoid overreliance. Organisations should incorporate these skills into ongoing workforce development to support effective human-AI teaming.

Organisations should also ensure AI systems used for cyber security, and AI systems used operationally, are protected from adversarial techniques such as prompt injection, model evasion and model extraction, in line with *Guidelines for secure AI system development*, National Institute of Standards and Technology's (NIST) *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* and MITRE ATLAS.

Human oversight

AI should support cyber defenders, not replace human judgment, particularly where decisions could affect high consequence cyber or safety environments. As AI systems become more capable and begin recommending or taking actions, strong human oversight is essential.

AI use introduces risks that can degrade reliability if not properly managed. These risks include hallucinated or misleading outputs, adversarial manipulation (such as prompt injection or model evasion), overreliance on AI, oversight fatigue and operational pressures related to scale, availability or cost.

While AI can significantly increase speed and scale, organisations should verify its outputs against evidence and context. Ongoing monitoring and human oversight help ensure that AI-enabled acceleration improves accuracy and reliability, rather than introducing new errors.

Organisations should:

- ensure AI supports, rather than replaces, human accountability for cyber security outcomes

- define clear limits on AI-driven automation and require human review and approval for actions with significant security, operational or safety impacts, particularly changes to access controls, security policies, critical systems and services, with accountability for high-consequence actions remaining with authorised personnel
- limit autonomous actions to those that are narrowly scoped, preapproved and reversible
- verify AI-generated outputs against evidence and context before action, with particular attention to detecting hallucinations or misleading outputs
- validate AI-generated recommendations affecting safety, availability, segmentation, recovery or process integrity in OT environments with qualified OT, engineering and safety personnel before implementation
- continuously monitor and review AI performance and behaviour
- include AI-related failures, compromise and hallucinations in incident response plans across IT environments.

The use of agentic AI introduces additional risk because these systems can initiate and carry out actions with reduced human involvement. Organisations should approach the careful adoption of agentic AI services, ensuring their use is tightly governed, clearly limited and supported by strong human oversight.

System protection and sandboxing

Organisations using AI tools for cyber security should design and deploy them so failures, misuse or manipulation cannot directly harm systems, data or operations. Controls should limit what AI systems can access, influence and execute regardless of how or where they are integrated.

Organisations should:

- deploy AI tools securely by default, with conservative configurations and minimal privileges
- constrain AI behaviour and authority through sandboxing, scope restriction and execution limits
- apply technical guardrails that prevent AI systems from performing high impact or irreversible actions without appropriate control
- make AI-assisted actions traceable and reviewable, supporting oversight, investigation and rollback where required
- protect AI components and interfaces against misuse, manipulation or unauthorised access.

Secure system integration

Organisations integrating AI tools into existing systems, workflows and operational processes should ensure that they improve security without weakening organisational control. Manage AI

capabilities as part of the enterprise environment with architectural, security and governance standards consistent with other systems.

Organisations should:

- integrate AI tools using approved architectures, consistent with enterprise security and system design standards
- use controlled and auditable integration methods, such as purpose-built security platforms or secured API integrations
- avoid poorly governed integrations, particularly where AI systems can directly initiate actions in operational or high availability environments
- ensure AI-assisted outputs and actions remain visible within existing workflows and decision-making processes
- align AI integrations with operational processes, including change management, monitoring and incident response
- implement safeguards when ingesting logs and artefacts into other AI systems to manage prompt injection and untrusted input risks.

Governance

AI adoption should be deliberate, authorised and accountable, particularly where sensitive information or critical systems are involved. While AI can support analysis and decision making, accountability for outcomes should remain with the human cyber defender.

Organisations should:

- document and design AI use cases ensuring alignment with business objectives, cyber security priorities and organisational risk tolerance
- restrict access and modification rights to authorised user accounts, and limit administrator privileges to verified personnel
- define and enforce policies governing what data AI tools may access, how data may be used and where outputs may be transmitted
- prevent the sharing of sensitive information with unapproved AI tools, including unmanaged or consumer AI services
- ensure transparency and auditability, including the ability to explain inputs, authority and how outputs influenced decisions
- integrate AI oversight into existing governance processes, including risk acceptance, change management, assurance and incident response.

Supply chain

Adopting AI systems introduces unique supply chain risks that can threaten the cyber security of an organisation if not effectively managed. While the use of pretrained models, third-party datasets and external services can accelerate capability delivery, it may also introduce inherited compromise, hidden dependencies and reduced visibility over system behaviour.

Organisations should:

- maintain visibility across AI supply chain components, including models, data sources, third-party services and hosting arrangements supported by an AI bill of materials (AIBOM), SBOM or equivalent documentation
- assess AI-specific supply chain risks including data sensitivity, control of components and adversarial machine learning threats
- integrate AI supply chain risks into existing cyber security risk management and incident response processes
- conduct due diligence and define security expectations with AI vendors and suppliers, including vulnerability and incident reporting
- understand how organisational data is accessed, used, and handled across the AI supply chain to ensure alignment with data sovereignty requirements and to determine which types of AI solutions are most suitable
- build organisational awareness of AI supply chain risks, and ensure staff involved in AI development, deployment and operation understand relevant threats and secure practices.

For more information on AI supply-chain risks and mitigations, refer to:

- *Artificial intelligence and machine learning*
- The G7 Cybersecurity Working Group's *A Shared G7 Vision on Software Bill of Materials for Artificial Intelligence*
- NIST's *Securing AI Ecosystems: The Critical Role of AI Bills of Materials (AIBOM) in Mitigating Software Supply Chain Risks*.

Testing and assurance of AI tools

Organisations should test AI tools used for cyber security to confirm they operate as intended and produce outputs that are accurate, reliable and useful in practice. Vendor claims, demonstrations and limited testing are not enough to show that a tool will perform well in a real environment.

Testing needs to focus on how the tool behaves in the organisation's own environment. As systems, data and the threat environment change, organisations need to monitor whether the tool remains effective. If outputs become unreliable or the tool no longer meets operational needs, the organisation needs to review the tool's role and make any necessary changes.

Organisations should:

- test AI tools in their own environment prior to production use
- validate that AI outputs are accurate, relevant and aligned with observed evidence and context
- assess whether AI recommendations are actionable and support confident decision making by staff
- evaluate how AI tools perform under realistic conditions, including noisy data, incomplete information and malicious activity
- test how cyber defence activities continue when AI tools fail, are compromised, or produce unreliable outputs, including fallback and escalation arrangements
- monitor ongoing performance to identify degradation, drift or changes in accuracy as data and the threat environment evolves
- review the role and configuration of AI tools when outputs become unreliable or no longer meet operational needs.

AI Secure by Demand

Organisations should select, implement and operate AI systems with security built in from the outset, not added as an afterthought. This is the Secure by Demand principle, in support of the Secure by Design foundations, and it is relevant to the entire AI supply chain. It means that AI suppliers (and intermediaries) have a responsibility to develop products that are Secure by Design. Secure by Demand means that the customers (and intermediaries) should hold the suppliers to account for that responsibility by only purchasing Secure by Design products.

For more information, refer to ASD's:

- *Secure by Design foundations*
- *Choosing secure and verifiable technologies*.

Conclusion

AI offers a valuable opportunity to strengthen cyber defence by reducing reliance on manual processes, improving risk prioritisation, and enabling earlier detection and response to threats. When organisations integrate AI into fit for purpose security tools and aligned with established cyber security fundamentals, AI helps them adapt to an increasingly complex and fast-moving threat environment.

However, AI adoption in cyber security should be deliberate and well governed. Poorly implemented AI systems can increase organisational risk by creating new vulnerabilities and introducing unpredictable failure behaviours. Used thoughtfully, AI can strengthen cyber security; used carelessly, it can make organisations more vulnerable.

Further information

ASD resources

- Artificial intelligence
- *Deploying AI systems securely*
- *Principles for the secure integration of Artificial Intelligence in Operational Technology*
- *Managing cyber supply chains*
- *Choosing secure and verifiable technologies*
- Secure by Design

Other resources

- NCSC-UK's *Supporting AI adoption for UK cyber defence*
- European Telecommunication Standards Institute's *A world first: the ETSI global cybersecurity standard for AI*
- Canadian Centre for Cyber Security's *Frontier artificial intelligence*

Contact details

If you have any questions regarding this guidance you can write to us or call us on 1300 CYBER1 (1300 292 371).

Appendix A: Cyber security questions for AI vendors

Appendix A provides a set of questions to help organisations assess the security, reliability and maturity of AI capabilities offered by vendors. These questions are designed to distinguish engineered, production-ready solutions from those relying on automation, assumptions, or emerging features.

Measurable security outcomes

- What specific security outcomes does the AI measurably improve (such as Mean Time to Detect, Mean Time to Respond, false positives/negatives)?
    - Ask for quantified evidence based on operational use rather than architectural or marketing claims.
- How do you validate and test AI performance in real security operations, not just lab scenarios?
    - To see if the vendor undertakes continuous evaluation, red teaming, customer side validation and production metrics.
- Which parts of the AI capability are production grade today versus roadmap items?
    - This helps distinguish proven capability from forward looking or experimental features.

Scope, capability and limitations

- Is the AI constrained to predefined rules and workflows, or can it reason beyond engineered use cases?
    - This clarifies whether the product is a narrow task-specific tool or a more general analytical capability.
- How does the AI model handle novel or previously unseen attack techniques?
    - Look for explanations of generalisation, adaptation, or pattern learning rather than reliance on static signatures.

Human oversight and decision making

- How do you prevent automation bias among analysts using the AI?
    - Look for design features that encourage review, challenge and critical thinking rather than blind trust.

- Who holds accountability when the AI makes an error and how are errors identified, investigated and remediated?
    - This clarifies ownership, escalation paths and continuous improvement processes.

Transparency, explainability and auditability

- Can the AI explain its reasoning and recommendations in a way cyber defenders can audit and defend?
    - Transparency is critical for incident response, regulatory scrutiny and executive assurance.
- Can you demonstrate how AI inputs, decisions and outputs are logged and reconstructed after an incident?
    - This supports forensic review, compliance obligations and post incident learning.

Security of the AI system itself

- How is the AI protected against manipulation, poisoning, or malicious input?
    - Assess resilience against prompt injection, training data poisoning and feedback loop abuse.
- What access controls and monitoring protect the AI model, prompts and tuning mechanisms?
    - AI components should be treated as high value assets within the security architecture.

Data protection and sovereignty

- How is sensitive data handled, stored and protected during AI processing?
    - Ensure alignment with internal policies, regulatory requirements and customer data sovereignty obligations.
- What data is retained, reused, or used to retrain models and under what conditions?
    - This helps identify unintended data exposure or long-term privacy risks.
- Which jurisdictions process or store customer data, including through third-party AI services?
    - This is critical for organisations with cross border or regulated data constraints.

Integration and operational fit

- How does the AI enhanced product integrate with existing cyber security tools, workflows and data sources?
    - Clarify whether integration is native, API based or requires significant custom engineering.
- How does the AI integrate with existing Identity and Access Management, logging, monitoring, change management and incident response processes?
    - AI should strengthen existing security architecture, not operate as a standalone or opaque system.

Resilience, dependency and long-term risk

- What dependencies does the AI capability have on third-party models, services, or data sources?
    - This helps assess supply chain risk, resilience and exposure to upstream changes.
- How does the system behave if the AI component degrades, fails, or produces low confidence outputs?
    - This helps understand fallback mechanisms and whether security operations can continue safely.
- What controls exist to restrict, scope, or disable AI functionality if risk tolerance or operating conditions change?
    - This is essential for incident response, regulatory change, or shifts in organisational risk appetite.
- How portable are the AI capabilities if the organisation changes vendors or deployment models?
    - This helps identify lock-in risks, particularly where detections or workflows are proprietary.
- Does the vendor provide a machine readable SBOM and AIBOM for AI that meets the minimum elements?
    - This supports assessment of AI supply chain transparency and dependency risk.
- Can the vendor demonstrate clear model provenance and lineage throughout the AI lifecycle?
    - This supports assurance, traceability and incident investigation.

Appendix B: AI capability questions for cyber security vendors

Appendix B provides structured questions to help organisations evaluate the effectiveness, reliability and operational fit of AI capabilities in cyber security. The focus is on ensuring AI delivers measurable outcomes while remaining secure, transparent and aligned with organisational risk requirements.

Security outcomes and effectiveness

- Does the AI capability measurably improve how quickly or accurately security teams can make decisions?
    - AI that improves decision speed or quality provides operational value; AI that only adds information can increase workload and risk.
- Which security activities are demonstrably improved by the AI compared to existing non-AI approaches?
    - This helps determine whether AI meaningfully changes outcomes rather than replicating existing capability with additional complexity.

Scope, assumptions and constraints

- What assumptions does the AI make about data quality, system architecture or threat behaviour?
    - Understanding assumptions helps identify where outputs may become unreliable outside validated conditions.
- What technical controls define where the AI cannot operate or influence outcomes?
    - Hard constraints reduce the risk of misuse or unintended expansion beyond approved use cases.

Handling uncertainty and error

- How does the AI communicate uncertainty, low confidence or incomplete information?
    - Clear signalling of uncertainty is essential to avoid false confidence and inappropriate response actions.
- What safeguards limit the impact of incorrect or misleading AI outputs?
    - Containment mechanisms reduce the blast radius of errors when operating at scale or speed.

Transparency and auditability

- Can security teams trace how specific inputs contributed to a given AI recommendation?
  - Traceability enables analysts to validate, challenge and defend AI-supported decisions.
- Are AI actions and recommendations logged in a way that supports post incident review?
  - Audit ready logs are essential for forensic investigation and assurance.

Security of the AI system

- How is the AI protected from adversarial manipulation, including prompt injection and model evasion?
  - AI used for defence is itself a target and must be secured as part of the attack surface.
- What monitoring exists to detect misuse, degradation or abnormal AI behaviour?
  - Ongoing monitoring helps identify emerging risks before they affect operations.

Data use and protection

- What operational data does the AI access, and how is access minimised?
  - Limiting data access reduces exposure and supports least privilege principles.
- Is any customer data reused beyond its original purpose, including for model tuning?
  - Clear boundaries on data use reduce long-term privacy, confidentiality and legal risk.

Integration and operational fit

- How does the AI integrate into existing security workflows without creating parallel processes?
  - Tight integration preserves governance and reduces operational friction.
- How does the AI behave when integrated data sources become unavailable or degraded?
  - Predictable behaviour under degraded conditions supports safe continued operations.

Resilience and recovery

- Can AI components be rolled back or isolated without disrupting core security functions?
  - Rollback capability is essential after compromise, misconfiguration or model failure.
- How are AI systems tested for safe failure and recovery scenarios?

- Resilience testing ensures security operations remain reliable under stress.

Disclaimer

The material in this guide is of a general nature and should not be regarded as legal advice or relied on for assistance in any particular circumstance or emergency situation. In any important matter, you should seek appropriate independent professional advice in relation to your own circumstances.

The Commonwealth accepts no responsibility or liability for any damage, loss or expense incurred as a result of the reliance on information contained in this guide.

Copyright

© Commonwealth of Australia 2026

With the exception of the Coat of Arms and where otherwise stated, all material presented in this publication is provided under a Creative Commons Attribution 4.0 International license https://creativecommons.org/licenses/by/4.0/.

For the avoidance of doubt, this means this license only applies to material as set out in this document.



The details of the relevant license conditions are available on the Creative Commons website as is the full legal code for the CC BY 4.0 license https://creativecommons.org/licenses/by/4.0/legalcode.en.

Use of the Coat of Arms

The terms under which the Coat of Arms can be used are detailed on the Department of the Prime Minister and Cabinet website https://www.pmc.gov.au/resources/commonwealth-coat-arms-information-and-guidelines.

cyber.gov.au