principia ratio

ARTIFICIAL INTELLIGENCE (AI) · BOARD BRIEFING

Secure AI adoption.

What Australian Government guidance now expects of boards, as AI starts to act on systems and attackers use it to move faster.

Own both sides of the risk.

Limit what AI can reach and do.

Ask for evidence that the limits hold.

Fix the fundamentals faster.

A synthesis of Protective Security Policy Framework (PSPF) advice, Australian Cyber Security Centre (ACSC) guidance, Information Security Manual (ISM) controls and the Australian Signals Directorate (ASD) Essentials series.



PARLIAMENT HOUSE, CANBERRA

ACCOUNTABILITY

AI risk sits with the board.

The September 2026 ISM makes the board or executive committee accountable for AI. Guidance for directors asks them to oversee two risks at once.

> ‘The board of directors or executive committee is accountable for ensuring that artificial intelligence is secure, controllable, human-supervised and used in an ethical and accountable manner.’

ISM cyber security principle GOV-08, September 2026 [8]

RISK ONE · OUR USE OF AI

AI that sees our data and acts for us.

Data can leave through a supplier. Staff can rely on wrong answers. Agents can do more than anyone intended.

RISK TWO · ATTACKERS’ USE OF AI

AI that finds our weaknesses first.

Attackers can use AI to find and exploit flaws within hours, and to chain minor flaws into a major incident. [6]

The ACSC asks for one named executive accountable for both risks. [5]

THE GUIDANCE

Who sets what.

Home Affairs sets policy. ASD advises and sets controls that can be tested. The PSPF binds Australian Government entities; other boards can use all three as a benchmark.

POLICY · HOME AFFAIRS

PSPF

Allows OFFICIAL information in generative AI, subject to normal system authorisation and checks on the provider. Calls for cyber fundamentals against AI-enabled attacks.

GUIDANCE · ASD

ACSC

Explains how to adopt AI and agents carefully, how to defend against AI-enabled attacks and what boards should ask.

CONTROLS · ASD

ISM

Turns both into controls that can be tested. By our count it holds 30 AI controls, at least 24 of them issued since December 2025.

FROM CHAT TOOLS TO AGENTS AND AI-ENABLED ATTACKS, IN UNDER THREE YEARS

JAN 2024
ACSC guidance on using AI securely [3]

OCT 2025
PSPF permits generative AI for OFFICIAL information [1]

DEC 2025
ISM adds its first large set of AI controls [8]

MAY 2026
Agent adoption guidance [4]; PSPF on AI-enabled threats [2]

JUL 2026
Defending against AI-enabled attacks [5]

SEP 2026
ISM agent register and limits; board and agent guidance [6–8]

OCT 2026
Essentials series treats AI as an asset [9], [10]

Approve uses, not just tools.

An approved provider is where the decision starts. Each use still needs an owner, a limit on the data it may touch and someone who accepts the risk.

PSPF SAYS

Generative AI hosted by a Hosting Certification Framework provider, OpenAI or Anthropic needs no further assessment of foreign ownership, control or influence. Assess any other provider first.

The system still needs authorisation by an officer who accepts its residual risk. Staff training must cover classified information. [1]

ACSC SAYS

Know where data is stored, whether inputs retrain the model and what happens to data when the contract ends.

Train staff on what they may enter and how far to trust the output. Weigh reliance on AI vendors, including their foreign ownership, control or influence. [3], [6]

ISM CALLS FOR

A maintained AI usage policy. No training on organisational data without the data owner's informed, explicit consent.

Prompts and outputs are deleted when a chat is deleted. A documented split of security duties with each supplier. [8]

For each use, know its owner, its data limits and who accepted the risk.

Limit what agents can do.

Agents use tools and change records. Their risk depends less on the model than on what they can reach and what they may do without asking.

DRAFTS

Assistant

Drafts and summarises. A person decides what to do with the result.

Main control: data limits and checked outputs.

READS

Connected assistant

Reads organisational documents, mail and databases.

Main control: access by role; treat what it reads as untrusted.

ACTS

Agent

Calls tools, sends messages and changes records.

Main control: its own identity, least privilege and human approval.

Start small. Begin with low-risk, well-defined tasks. Widen access and autonomy only as controls prove themselves. [4], [7]

Apply the narrower permission. An agent may do only what both the user and the task allow. Nothing it reads may change its instructions or approvals. [8]

Register every agent. Each gets its own identity and an entry recording its owner, purpose, credentials, tools and data. [8]

People set the approval points. Designers or operators, not the agent, decide which actions need a person's approval. [4], [8]

Illustration. An agent that reads supplier invoices should have no way to change supplier bank details, whatever an invoice tells it to do.

Treat AI as an asset.

ASD's Essentials series, which replaces the Essential Eight, sets no separate framework for AI. Each AI system is placed in a security domain by what it reaches and what its loss would cost. [9]

AI SYSTEM	LIKELY DOMAIN
Agents and the accounts they use	Identity services
AI tools that can take administrative actions	Privileged environments
Public-facing AI services, interfaces and connectors	Boundary services
AI that supports core organisational outcomes	Critical processes and services
AI built into general business software	Business systems

Approve before switching on. Material AI features in cloud services need approval from an executive or their delegate, after a review of data, permissions, retention, logging, agents and connectors. [10]

Contain, do not just filter. Prompt injection cannot be completely prevented, so limit what a manipulated model can reach. [9]

Keep AI away from safety systems. In operational technology (OT), no AI platform should have a path that can change them. [9]

Plan for impersonation. Synthetic voice and video target help desks and payment approvers. Check unexpected requests through a separate channel. [9]

Ask which AI features were switched on this year, and who approved each one.

Ask for records, not reassurance.

Management should be able to show the board these six records for every AI system in use.

RECORD	WHAT IT SHOWS	BASIS
Use and agent register	Every use, agent and cloud AI feature, with its owner, purpose and access.	ISM-2134–2135; [10]
Supplier terms	Where data goes, how long it is kept, whether it trains models and who secures what.	[3]; ISM-1569, 2103
Test results before launch	How the system behaved when given misleading input, including instructions hidden in documents.	[3], [4]; ISM-1924, 2158
Action and approval logs	Every tool call, outside request and human approval, kept for investigation.	[7]; ISM-2159
Normal behaviour	Expected performance, volume and cost, with alerts when they change.	[3]; ISM-2089–2091, 2114
Stop and recovery plan	Who can halt an agent, revoke its access and reverse its changes.	[3], [4]

Our test: management can list every agent, stop any one and show who approved its last high-impact action.

Attackers can now move in hours.

The PSPF says government entities do not need the most advanced AI models to stay protected. The answer is the existing fundamentals, applied faster and more consistently. [2]

WHAT HAS CHANGED

The time from finding a flaw to exploiting it is shrinking from days to hours. [5], [6]

Attackers chain several minor flaws into one serious compromise.

Less skilled attackers gain advanced capability.

Internet-facing and unsupported systems are exposed first. [5]

The PSPF warns of a 'vulnerability storm' from known flaws, legacy systems and weak basic security. [2]

NOW

Close the doors.

Lock systems to approved settings. Patch and verify exposed weaknesses quickly.

SHORT TERM

Limit the damage.

Replace or isolate legacy systems. Tighten identity and access, including for agents. Rehearse incident plans.

MEDIUM TERM

Use AI in defence.

Apply AI to monitoring, vulnerability testing and code review, under human supervision.

LONGER TERM

Modernise.

Rebuild systems so they are secure by design and by default.

For Australian Government entities, the PSPF also requires Essential Eight Maturity Level Two for patching and for hardening user applications. [2] ASD's Essentials series has since replaced the Essential Eight, but a target the PSPF requires still stands.

DECISIONS FOR THE BOARD

Five decisions to make now.

The ACSC lists the first four as immediate actions, three of them for the board or executive committee to decide. The fifth follows from the PSPF, the ISM and the Essentials.

	DECISION	WHAT IT MEANS IN PRACTICE	BASIS
1	**Name one executive accountable for AI risk.**	One person answers for both our use of AI and attackers' use of it.	[5]; GOV-08 [8]
2	**Accept brief outages to patch faster.**	Internet-facing services may go down briefly for urgent fixes, more frequent patching or a faulty patch.	[5]
3	**Lower tolerance for unsupported systems.**	Remove or replace exposed systems the vendor no longer supports, with extra monitoring until then.	[2], [5]
4	**Fund the work.**	Allocate people and money for the fundamentals and for AI controls.	[5], [6]
5	**Set the limits of AI use.**	Approve which uses, agents and cloud AI features may run, and which actions always need a person.	[1], [4], [10]; ISM-2113, 2134

Ask for regular reports on the agent register, exposed weaknesses and unsupported technology.

QUESTIONS FOR MANAGEMENT

What directors should ask.

Drawn from ASD's guidance for directors and its questions for executives on agents. [6], [7]

Our use of AI

1. Which AI uses and agents are approved, and who owns each one?
2. What data, systems and tools can each agent reach? Which of its actions need human approval?
3. Could a document, email or web page instruct an agent to act? How do we know it cannot?
4. If an agent were compromised, what is the worst outcome, and how fast could we stop it?
5. Do we understand our AI vendors' security and their foreign ownership, control or influence?

Attackers' use of AI

1. Which parts of the business are most exposed if AI is used to find and exploit our weaknesses?
2. Which known weaknesses become far more dangerous if exploited at machine speed?
3. What legacy technology are we carrying, and when will it go?
4. If attacks moved from days to hours, could we still detect and respond in time?
5. Have our incident and continuity plans been tested against AI-enabled attacks?

REFERENCE NOTES

Sources and supporting controls.

Prepared for discussion. Apply each requirement according to the organisation's obligations and the AI it runs.

[1] Department of Home Affairs, PSPF Policy Advisory 001-2025: OFFICIAL information use with generative artificial intelligence, 7 October 2025.

[2] Department of Home Affairs, PSPF Policy Advisory 001-2026: Cyber security readiness in the frontier AI era, 27 May 2026.

[3] ACSC, Engaging with artificial intelligence, 24 January 2024.

[4] ACSC and international partners, Careful adoption of agentic AI services, 1 May 2026.

[5] ACSC, Defending against AI-enabled cyber attacks, 6 July 2026.

[6] ASD with the Australian Institute of Company Directors, Frontier AI cyber threat considerations for boards of directors, 8 September 2026.

[7] ACSC, Agentic AI harnesses, 11 September 2026.

[8] ASD, Information Security Manual, September 2026 edition: system security plan annex template.

[9] ASD, AI considerations within the Essentials, ASD Essentials overview, October 2026.

[10] ASD, ASD Essentials for cloud, chapter, October 2026.

THEME	ISM REFERENCES
Board accountability and usage policy	GOV-08; 2074
Data use, deletion and supplier duties	2103; 2123; 1569
Agent identity and register	2133–2135
Agent tools, permissions and untrusted content	2092–2093; 2156–2158
Human approval and expected behaviour	2113; 2114
Logging, monitoring and usage limits	2159; 2089–2091
Misleading input and output filtering	1924; 2094
Model integrity and documentation	2072; 2084–2088
Sensitive and classified data; private devices	2112; 2095
AI in cyber defence	2117; 2119; 2122

Source material © Commonwealth of Australia. Summarised and adapted by Principia Ratio; no government endorsement implied.