Securing AI in Australia’s Public Sector

Cyber Security Readiness for the AI Era

Rachelle Koster

rachelle.koster@resonantex.com.au

March 2026

A note on scope

This paper focuses on cyber security readiness as a precondition for responsible AI deployment. It does not address the broader responsible AI governance landscape (privacy, administrative law implications of automated decision-making, and algorithmic transparency) which require parallel attention but distinct analysis.

The step change driving this analysis is the arrival of generative AI from late 2022 onward. Australian government agencies have used predictive machine learning for years: fraud detection, visa risk scoring, revenue matching. Generative AI introduces fundamentally different security challenges: natural language input, novel outputs, and attack surfaces (prompt injection, data poisoning, model supply chain compromise) that sit outside existing control frameworks. Where this paper refers to “AI” without qualification, it means generative AI and foundation-model-based systems. Predictive ML is referenced where evidence specifically addresses it.

This analysis focuses on civilian Commonwealth government agencies and the federal policy architecture. Defence and Intelligence agencies, which operate under separate security frameworks, are not addressed. Every state and territory has published at least guidance-level AI policy, with maturity varying widely [33]. State agencies do however take some of their guidance from ACSC so there are takeaways within this paper which are relevant at a State or Territory level.

> **About this paper**
>
> This paper was researched and analysed with the support of Claude (Anthropic) as an AI research and analysis partner. Quality review was conducted using a proprietary sovereign AI solution. References also checked with Perplexity. The paper has been independently reviewed end to end by two public sector cyber security experts.
>
> All analysis draws on publicly available information. Sources were verified as accessible at the time of publication (March 2026).

1. Executive Summary

This paper examines whether Australia's public sector cyber security foundations are ready for the AI adoption the Government has committed to. It finds that the security reality does not match the AI policy ambition, and provides **seventeen recommendations** to uplift cyber security readiness within existing institutional structures.

Three years after the Essential Eight became mandatory, and following a 2023 tightening of the maturity model, **78% of Commonwealth entities have not yet met the baseline** overall across all eight mitigation strategies [1]. Independent audits across a decade document a consistent pattern: privileged access failures, incomplete asset inventories, untested incident response, and self-assessments that overstate reality. Into this environment, AI is arriving at pace. The number of Commonwealth entities using AI doubled from 27 to 56 in a single year [21], Microsoft Copilot trials ran across agencies in 2024, and GovAI launched in July 2025. GovAI Chat is expected from mid-2026, agencies are expected to appoint Chief AI Officers in 2026, the first new mandatory AI policy requirements from June 2026 with remaining requirements from December 2026 [31]. The ambition is necessary but the security foundations are not yet ready to support it.

Two challenges run in parallel. The first is amplification: existing cyber security gaps will be widened by AI adoption. Unresolved access failures become AI data exposure pathways. Low data governance maturity compounds the problem: agencies that cannot govern their data cannot secure AI systems that transform it. The second is genuinely new: prompt injection, model poisoning, adversarial attacks, and AI supply chain compromise sit outside existing frameworks. Both must be addressed together.

The policy architecture has improved markedly. PSPF 2025 reforms, Cyber Security Act 2024, Modern Defensible Architecture (MDA), and two consecutive material ISM quarterly updates (December 2025 and March 2026) adding AI-specific controls covering application development, governance accountability, and device-level restrictions. ASD is actively engaging with the IRAP assessor community and industry stakeholders on how to build out AI security in the ISM. That is the right approach. The structural challenge is pace: quarterly ISM updates, stakeholder consultation cycles, and agency budget processes operate on a cadence fundamentally slower than AI adoption. The gap is in implementation capacity: people, funding, and institutional willpower, and in the deployment-layer controls that remain unaddressed.

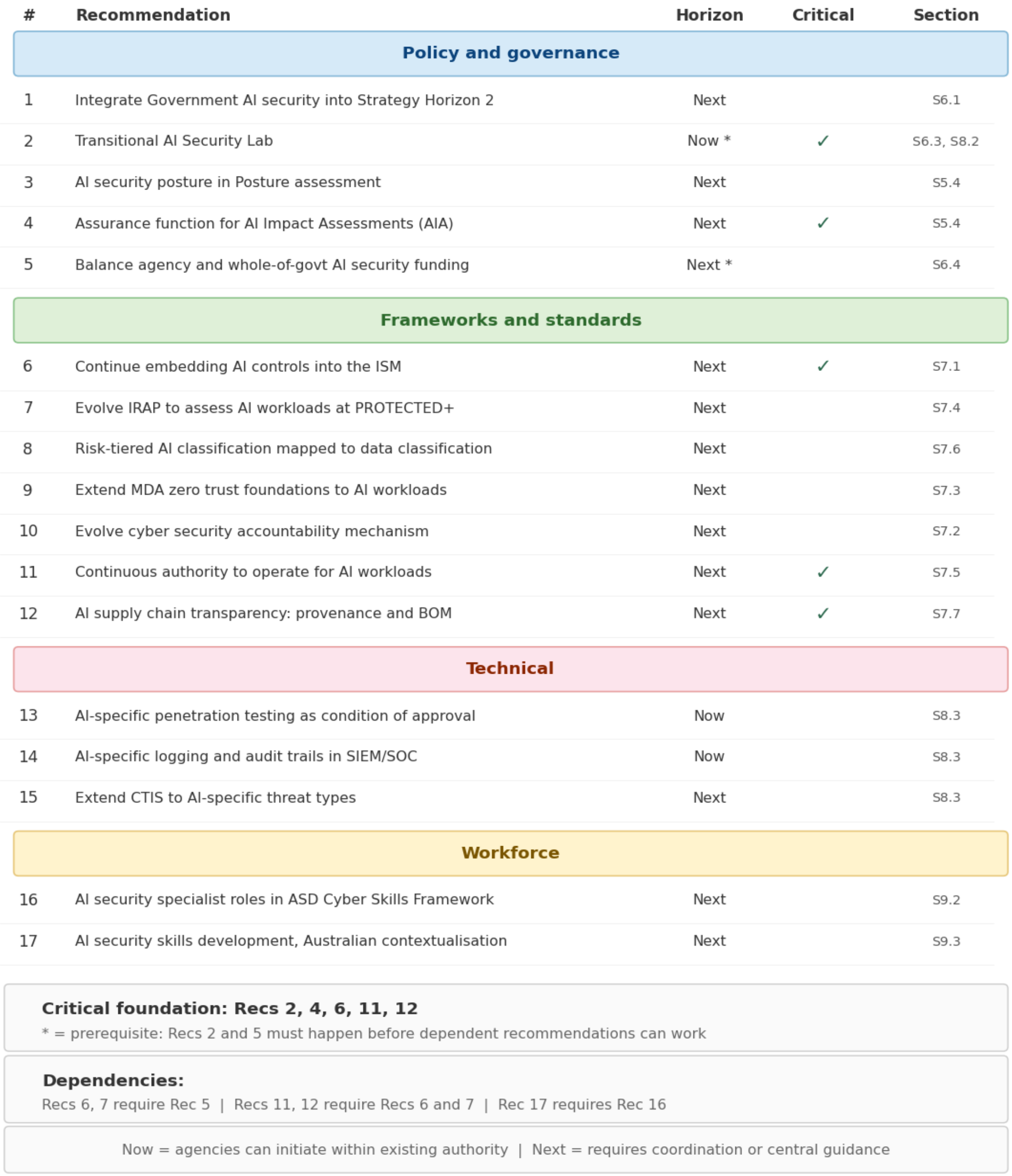
Seventeen recommendations outlined in this paper address that gap. **Five are critical**: a transitional AI Security Lab (Rec 2); independent assurance for AI governance before self-assessment optimism repeats (Rec 4); systematising the ISM's new AI controls across the deployment layer (Rec 6); continuous authority to operate for AI workloads (Rec 11); and AI supply chain transparency (Rec 12). Each is tagged by implementation horizon: what agencies can initiate now and what requires central coordination.

AI adoption is accelerating into production. If we tackle AI security with the same pace, budget cycles, and accountability mechanisms that have delivered 78% non-compliance with the current baseline, the security posture will not improve. It will likely deteriorate given the speed and scale of AI threats.

Table of Contents

2. Consolidated Recommendations

Seventeen recommendations are organised across four categories and tagged by implementation horizon: **Now** (agencies can initiate within existing authority), **Next** (requires coordination or central guidance). Five are critical: a transitional AI Security Lab (Rec 2); independent assurance for AI governance (Rec 4); systematising the ISM's new AI controls across the deployment layer (Rec 6); continuous authority to operate for AI workloads (Rec 11); and AI supply chain transparency (Rec 12). Each recommendation is developed in the analytical section indicated.



#	Recommendation	Horizon	Critical	Section
	Policy and governance			
1	Integrate Government AI security into Strategy Horizon 2	Next		S6.1
2	Transitional AI Security Lab	Now *	✓	S6.3, S8.2
3	AI security posture in Posture assessment	Next		S5.4
4	Assurance function for AI Impact Assessments (AIA)	Next	✓	S5.4
5	Balance agency and whole-of-govt AI security funding	Next *		S6.4
	Frameworks and standards			
6	Continue embedding AI controls into the ISM	Next	✓	S7.1
7	Evolve IRAP to assess AI workloads at PROTECTED+	Next		S7.4
8	Risk-tiered AI classification mapped to data classification	Next		S7.6
9	Extend MDA zero trust foundations to AI workloads	Next		S7.3
10	Evolve cyber security accountability mechanism	Next		S7.2
11	Continuous authority to operate for AI workloads	Next	✓	S7.5
12	AI supply chain transparency: provenance and BOM	Next	✓	S7.7
	Technical			
13	AI-specific penetration testing as condition of approval	Now		S8.3
14	AI-specific logging and audit trails in SIEM/SOC	Now		S8.3
15	Extend CTIS to AI-specific threat types	Next		S8.3
	Workforce			
16	AI security specialist roles in ASD Cyber Skills Framework	Next		S9.2
17	AI security skills development, Australian contextualisation	Next		S9.3

Critical foundation: Recs 2, 4, 6, 11, 12

* = prerequisite: Recs 2 and 5 must happen before dependent recommendations can work

Dependencies:

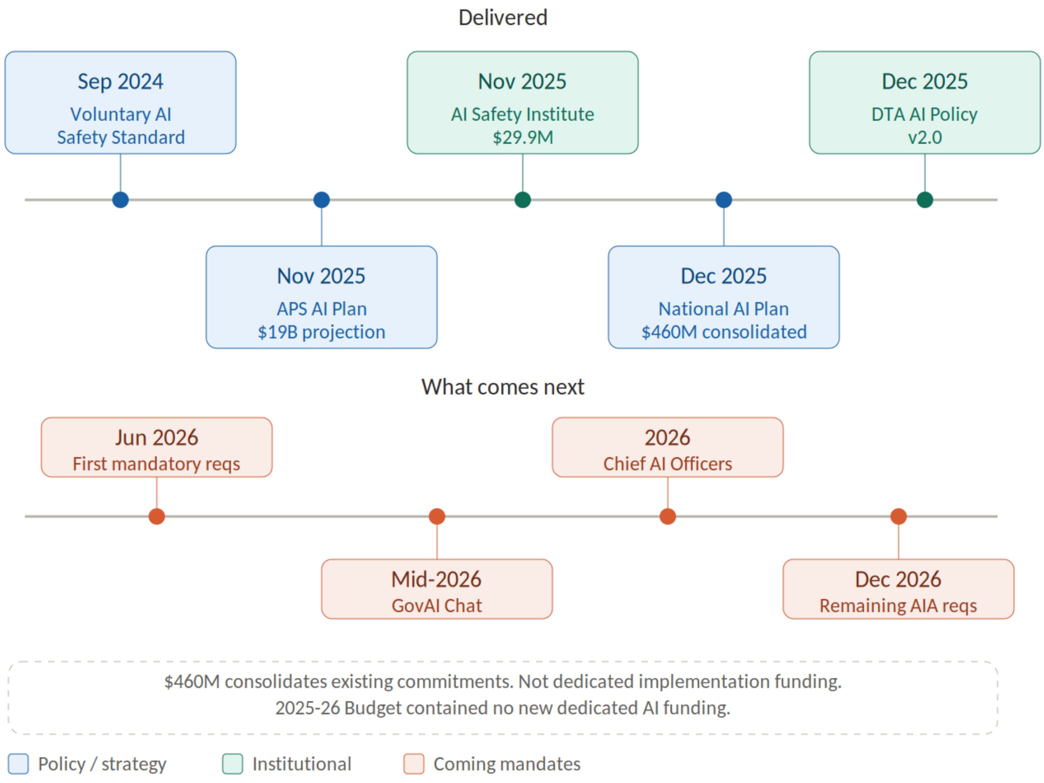
Recs 6, 7 require Rec 5 | Recs 11, 12 require Recs 6 and 7 | Rec 17 requires Rec 16

Now = agencies can initiate within existing authority | Next = requires coordination or central guidance

Recommendation matrix

3. Public Sector AI Policy & Program

In fewer than two years, the Australian Government moved from aspirational AI principles to binding mandates backed by broader AI-related commitments totalling over $460 million [35]. GovAI Chat is expected from mid-2026, agencies are expected to appoint Chief AI Officers in 2026, the first new mandatory AI policy requirements from June 2026 with remaining requirements from December 2026. The governance arc substantially advances within twelve months.



Federal AI Policy Timeline

The headline funding deserves scrutiny. The $460 million consolidates existing commitments: $362 million for research grants, $47 million for the Next Generation Graduates Program. This is not dedicated implementation funding for agency security uplift. The 2025–26 Budget contained no new dedicated AI funding [117]. The APS AI Plan estimates AI adoption could deliver $19 billion in annual value by 2030 [33]. This is an estimate, not an investment commitment. The money is concentrated in research pipelines. Outside GovAI Chat ($225.2 million), it is not flowing into operational deployment infrastructure or security foundations.

The sequencing warrants attention. GovAI Chat is expected from mid-2026; the first new mandatory AI policy requirements take effect from June 2026, with remaining requirements from December 2026 [31]. The governance instruments are arriving, and ASD has been building AI-specific controls into the ISM through two consecutive material quarterly updates: December 2025 (application development) and March 2026 (governance and device-level restrictions) [131]. Section 7 details what these updates cover and what remains unaddressed. ASD is engaging with IRAP assessors and industry stakeholders on the ISM's AI security roadmap. That is the right approach, producing better controls through consultation. The structural challenge is pace: quarterly ISM updates, stakeholder consultation, and agency budget cycles operate on a cadence fundamentally slower than AI adoption. An agency completing an AI Impact Assessment will still find no explicit ISM controls for model provenance, prompt injection defence, retrieval-augmented generation (RAG) pipeline security, and AI supply chain assurance. These are the deployment-layer gaps that matter most as AI moves from experimentation to production. DTA's March 2026 guidance on scaling AI from proof of concept to production correctly identifies the security instruments agencies should comply with (ISM, PSPF) [132]; the challenge is that most agencies have not met the pre-AI baselines those instruments require. In July 2025, DTA published the Australian Government AI Technical Standard, a substantial contribution to the governance landscape establishing 41 statements and 148 criteria across the full AI lifecycle [152]. This is where Recommendation 2, a transitional AI Security Lab, earns its place.

The argument is not that agencies should wait for perfect security before proceeding with AI. It is that agencies need a higher threshold for high-impact AI use cases, independent assurance for production deployments processing sensitive data, and tighter controls where AI interacts with sensitive government data. Getting the security foundations right enables AI adoption; deferring them guarantees that the amplification effect documented in Section 5 will compound.

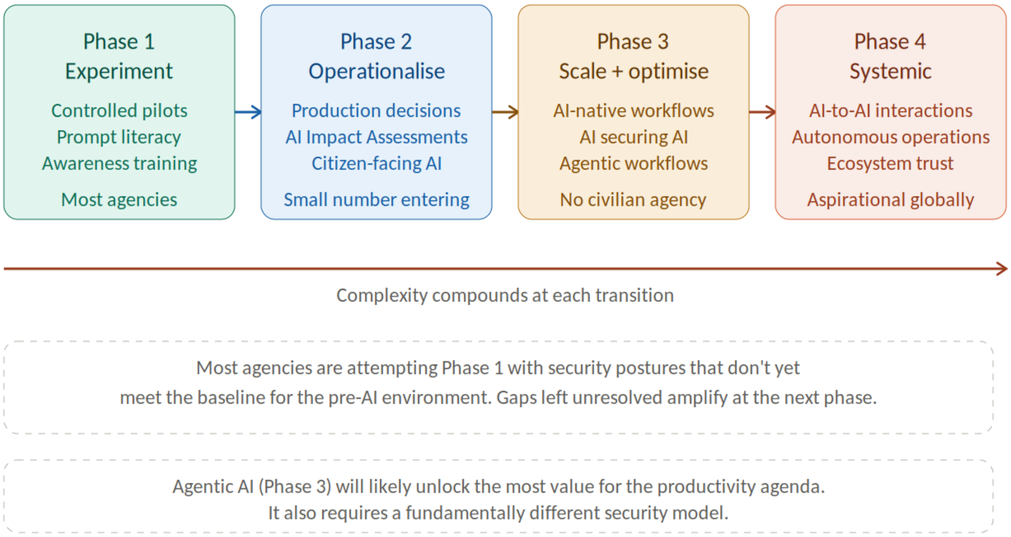
> **GovAI Chat: the boundary security question**
>
> The security question with GovAI Chat is not whether the platform itself is secure. Finance and ASD are designing it with appropriate security controls, and the phased rollout is the right approach [33].
>
> The harder question is what happens at the boundaries, where a centralised platform operating across more than one hundred agencies meets diverse security environments. Identity is the clearest example: every agency runs its own identity and access management at a different maturity level. When agentic AI enters the picture (non-human identities making decisions across agency boundaries), the complexity escalates beyond what current architectures were designed for.
>
> The likely architecture will see agencies building their own RAG pipelines connecting GovAI Chat to their document holdings, case management systems, and operational data. This keeps data sovereignty with agencies. That is the right approach. Following existing Commonwealth shared-service practice, accountability will follow the data: agency CISOs own their boundary security, and the platform operator sets minimum standards. For this model to work, agencies connecting to the platform need the controls in place before they connect: prompt-layer protection, RAG pipeline access controls, data classification enforcement at the retrieval layer, and logging that captures what was queried, what was retrieved, and from which data holdings.
>
> The detection and response challenge illustrates why some of this is better solved centrally. Consider an incident at 2am where a prompt, due to excessive or misconfigured cross agency privileges, retrieves sensitive data across an agency boundary: if the platform operator has continuous monitoring but the data-owning agency runs business-hours security operations with on-call support, who detects it? Who investigates? The 74% of government breaches that take more than 30 days to identify [52] will not improve when AI operates continuously across environments with inconsistent monitoring. Shared detection capability at the platform boundary layer, covering cross-agency prompt activity, anomalous retrieval patterns, and data access across agency holdings, is a stronger model than 190 agencies each independently monitoring their own GovAI boundary. This is one area where whole-of-government capability makes more operational and economic sense than distributed agency implementations.
>
> The Cyber Hubs experience is instructive. The shared-service logic was sound; operational interfacing with diverse agency security postures was where things broke down. GovAI Chat faces the same structural challenge, with the added complexity that AI systems process data faster, at greater scale,

> and with less visibility than conventional shared services. Independent assurance of agency boundary implementations (not self-assessed) is essential to counter the optimism bias documented throughout this paper.
>
> Many of the recommendations in this paper converge on making this model work: AI-specific logging (Rec 14), detection capability at AI-era operating tempo, independent assurance for AI Impact Assessments (Rec 4), the workforce to implement and assess these controls (Rec 16), and the funding to sustain them (Rec 5). GovAI Chat is the near-term forcing function. Getting this right would set the pattern for every shared AI service that follows.

4. The Journey Ahead for Public Sector AI

4.1 A four-phase adoption model

The government's AI timeline means agencies will move through a predictable adoption journey as outlined below. The critical insight is not where individual agencies sit today. Most are at Phase 1 with security postures that do not meet the pre-AI baseline [1]. It is that gaps left unresolved at one phase become amplified risks at the next.



Most agencies are attempting Phase 1 with security postures that don't yet meet the baseline for the pre-AI environment. Gaps left unresolved amplify at the next phase.

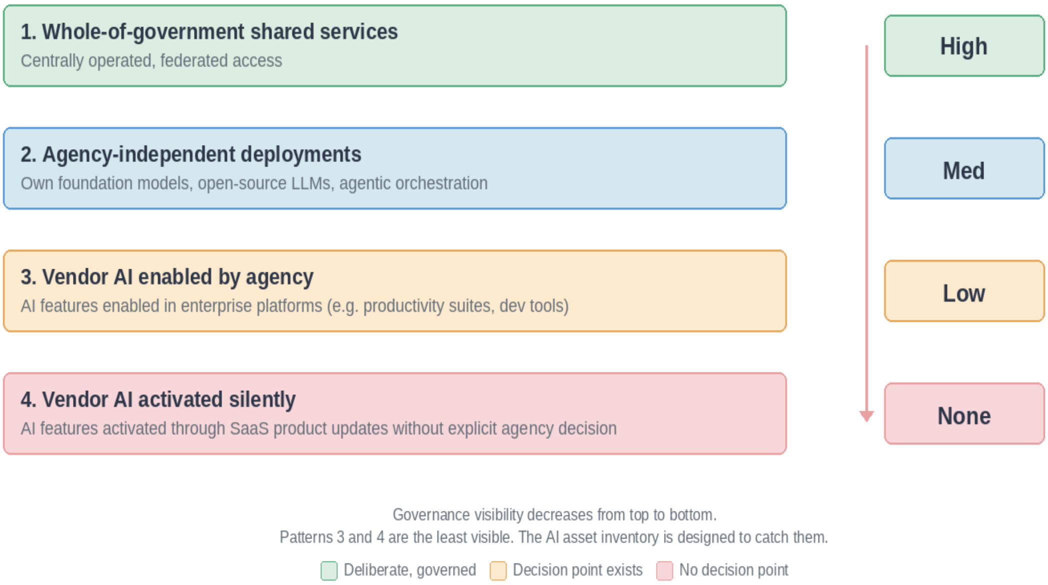
Agentic AI (Phase 3) will likely unlock the most value for the productivity agenda. It also requires a fundamentally different security model.

AI Adoption Journey

The security implications at each phase transition are examined in the sections that follow.

4.2 Four deployment patterns

GovAI Chat will not be the only AI used in the public sector. Government agencies will likely use a variety of patterns within their vendor ecosystem, with the larger agencies potentially having all four.

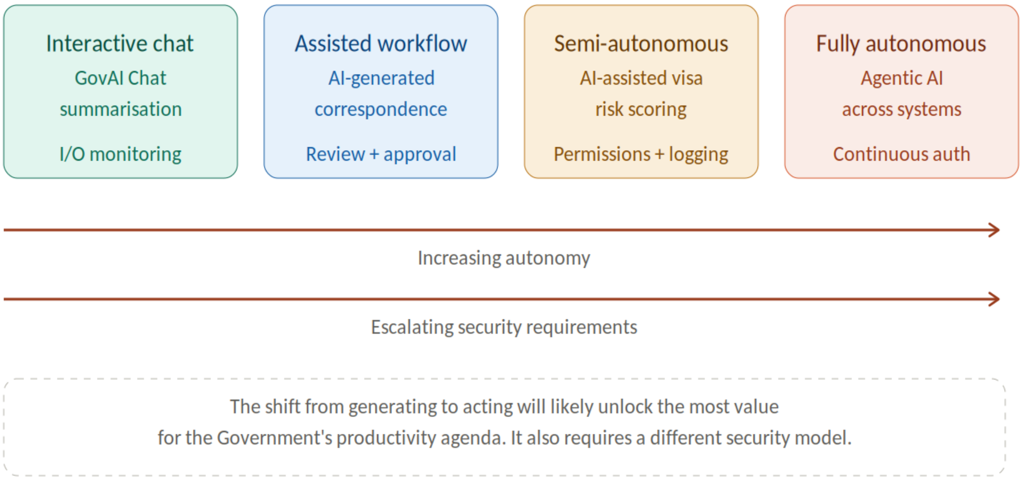

1. Whole-of-government shared services
Centrally operated, federated access
2. Agency-independent deployments
Own foundation models, open-source LLMs, agentic orchestration
3. Vendor AI enabled by agency
AI features enabled in enterprise platforms (e.g. productivity suites, dev tools)
4. Vendor AI activated silently
AI features activated through SaaS product updates without explicit agency decision
High
Med
Low
None
Governance visibility decreases from top to bottom.
Patterns 3 and 4 are the least visible. The AI asset inventory is designed to catch them.
🟩 Deliberate, governed
🟧 Decision point exists
🟥 No decision point

Four deployment patterns

Governance visibility decreases sharply from Pattern 1 to Pattern 4. Patterns 1 and 2 involve deliberate decisions that trigger governance processes. In Pattern 2, agencies are deploying their own foundation models, open-source LLMs, or agentic orchestration layers; these are independent architectural choices with full security ownership and no shared baseline. Pattern 3 involves a decision point, but the agency does not control the model, the training data, or the vendor's security architecture. The EchoLeak vulnerability (CVE-2025-32711) demonstrated zero-click data exfiltration from Microsoft 365 Copilot. Agencies deploying Copilot within M365 tenants are deploying the same platform this vulnerability targeted. Pattern 4 is the least visible: AI features activate through routine product updates without explicit agency decision and without triggering governance processes. Combined with the asset inventory blindness documented in Section 5, agencies may not know AI is operating. The AI asset inventory action in Section 8 is designed to catch Patterns 3 and 4.

4.3 Agentic AI: a different security model

AI in government will operate across a spectrum from interactive chat to fully autonomous agents. The security model shifts fundamentally when AI moves from generating responses to taking actions, and that shift will likely unlock the most value for the Government's productivity agenda.



AI autonomy spectrum

Agents calling APIs, writing files, and querying databases need a permissions model most agency environments do not yet have. Early 2026 saw rapid advances in device-level AI agents. MITRE’s OpenClaw research [148] demonstrated how autonomous agents on personal devices can be exploited through prompt injection, tool invocation, and configuration manipulation, and commercial tools with similar autonomous capabilities are proliferating. This trend poses direct challenges to existing endpoint security, data loss prevention, and BYOD programs where government staff access agency systems from personal devices. ISM-2095, introduced in the March 2026 update, responds to this risk: it explicitly restricts “unapproved artificial intelligence agents” on personal devices accessing OFFICIAL: Sensitive or PROTECTED systems [131]. The control addresses the immediate device-level threat, but as agentic deployments scale into enterprise workflows, the permissions and monitoring requirements will extend well beyond device policy. Recommendations 14 (logging), 11 (continuous authority to operate), and 9 (MDA zero trust) are designed to enable agentic deployments in agencies as they arrive.

AI agent security is a rapidly evolving field. An emerging agent integration stack is taking shape: the Model Context Protocol (MCP) for connecting agents to tools and data sources, protocols like Google’s Agent-to-Agent (A2A) for inter-agent coordination, and orchestration frameworks for workflow control and state management. Each layer introduces distinct security challenges. MCP illustrates the pattern: the specification includes OAuth 2.1 authorization as an optional capability (added in the March 2025 revision), but security researchers scanning the internet found 1,862 exposed MCP servers, with all 119 manually verified samples responding to unauthenticated requests [149]. The protocol mandates neither enterprise audit logging nor server identity verification. Agent-to-agent protocols and orchestration layers are at an even earlier stage of security development. As agencies deploy AI agents that connect to enterprise systems, coordinate with other agents, and execute multi-step workflows, the identity, logging, and supply chain gaps documented in this paper become immediate operational risks. This is one of the reasons why the ISM will need to continue evolving its AI controls as outlined in Section 7.1. The AI Technical Standard, published in July 2025 before agentic AI moved from concept to widespread adoption, does not address AI agents, tool use, function calling, or agent orchestration [152]. This is not a criticism of DTA. It illustrates precisely the pace challenge this paper documents: government standards face the same cadence constraints as the ISM in keeping up with a technology that evolves faster than any consultation-led process can follow.

4.4 The AI security threat environment

The introduction of AI into the public sector creates new attack surfaces and vulnerabilities. MITRE ATLAS catalogues 15 tactics, 66 techniques, and 33 real-world case studies specific to AI (as of October 2025) [80]. Approximately 70% of ATLAS mitigations map to existing security controls [115]. Nation-state actors are already targeting AI systems processing government data, and AI-specific attack techniques, including prompt injection, data poisoning, and AI supply chain compromise, are documented in active use. The defence relies on getting fundamentals right, then building AI-specific capabilities on top. A summary of the current AI threat landscape is at Appendix A.

4.5 AI-enabled security as part of the solution

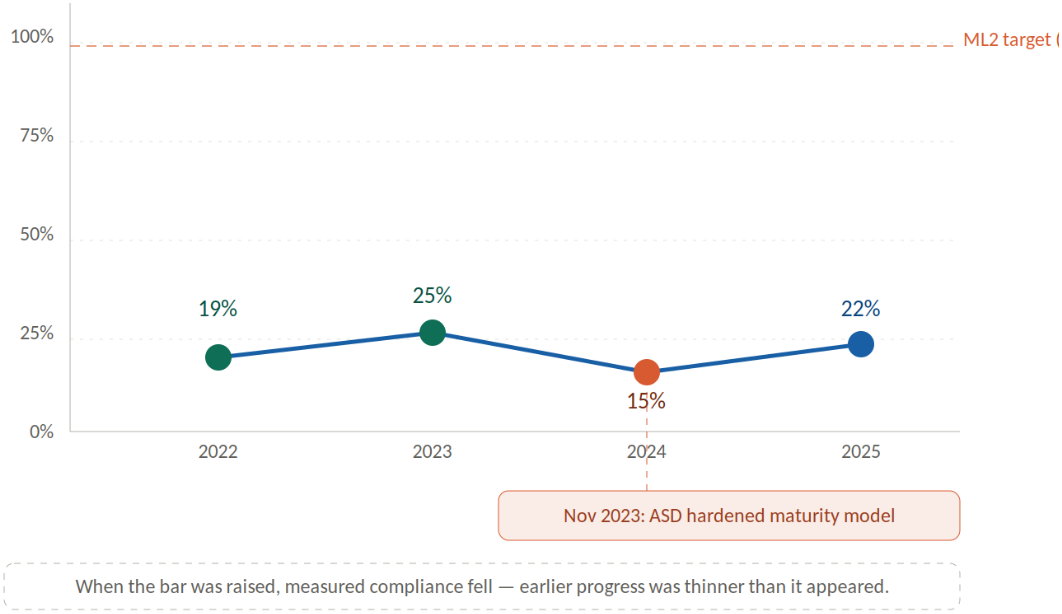
AI is of course part of the solution to address these security challenges. Security platforms are embedding AI capabilities across every major domain and new AI-native security platforms will be developed. For agencies, AI adoption makes platform modernisation unavoidable, and there should be alignment with other cyber security requirements such as MDA and SSE/SASE. However, this fundamentally remains a budget and policy challenge to give agencies the ability to procure and implement new platforms, not a technology availability problem.

5. The Current Cyber Security Posture

5.1 What the audits reveal

The Essential Eight is the most visible measure of Commonwealth cyber security because it gets publicly audited. It is also a narrow one: eight prioritised controls drawn from ASD’s broader set of 37 mitigation strategies [9], covering application control, patching, macro settings, user application hardening, administrative privileges, MFA, and backups. These eight address the attack patterns that account for the majority of conventional intrusions, and they are the only domain with a mandatory maturity threshold.

Three years after the Essential Eight became mandatory at Maturity Level 2, and following a 2023 tightening of the maturity model, 78% of Commonwealth entities have not yet met that baseline overall across all eight mitigation strategies [1]. Compliance moved from 19% (2022) to 25% (2023), dropped to 15% (2024), then partially recovered to 22% (2025) [1][2][3]. The 2024 drop resulted from ASD hardening the maturity model. Earlier progress was thinner than it appeared.



Nov 2023: ASD hardened maturity model

When the bar was raised, measured compliance fell — earlier progress was thinner than it appeared.

E8 ML2 Compliance Trajectory

The ANAO has tabled seven major cyber security audits since 2013–14. Across the first four, only 4 of 14 entities complied with mandatory controls [19]. The ANAO's conclusion: previous audits have not found significant improvement over time, and this pattern continues [18].

5.2 Beyond the Essential Eight: the environment AI lands in

Essential Eight compliance is the most audited measure, but it is one indicator within a broader landscape of weakness that directly affects AI readiness.

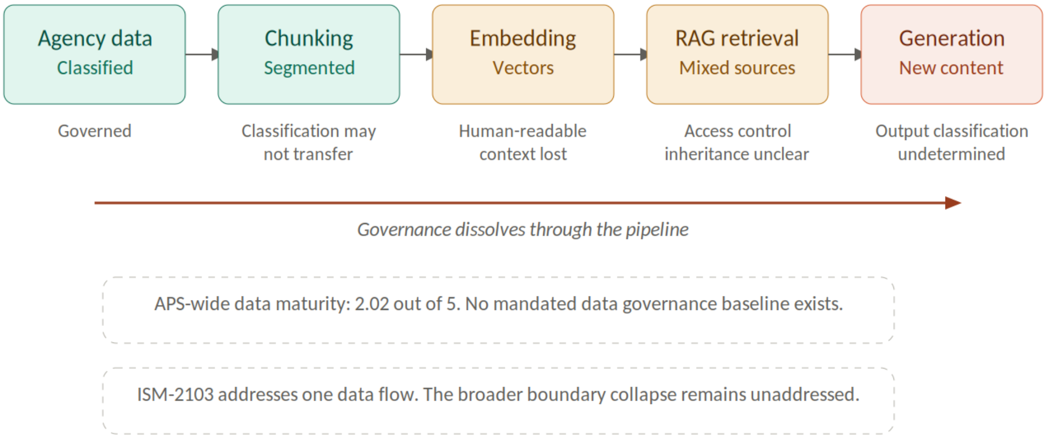
The structural barriers are interconnected. Fifty-nine per cent of entities report that legacy technologies impede Essential Eight implementation [1], with more than 70% still reliant on legacy systems [101] and Microsoft assess that only 10% of government IT spending is going to public cloud [101]. Cloud adoption is broad (89% of entities use cloud arrangements [21]), but AI workloads will run on the same infrastructure where these assurance gaps exist. Supply chain risk is accelerating. ANAO found contracts with providers often did not detail PSPF requirements [20], and AI deployment will dramatically expand the supply chain attack surface through model providers, training data pipelines, and API integrations.

Workforce shortages sit underneath all of these problems. More than half of APS agencies report critical cyber security skills shortages [69], with a projected digital workforce shortfall exceeding 8,000 by 2030 [70]. AI adoption adds new skill requirements to an already insufficient base. None of these gaps exists in isolation. AI lands in the middle of all of them simultaneously.

Data governance: the compounding factor

The APS-wide average data maturity score is 2.02 out of 5, rated "developing" on the Department of Finance Data Maturity Assessment Tool (DMAT) scale: agencies have started initiatives but lack systematic, consistent data governance practices. “Data Quality, Reference & Metadata” scored lowest [137]. The ANAO confirmed “fundamental deficiencies” across entities [138]. Unlike cyber security, no mandated data governance baseline exists.

This matters because AI fundamentally transforms how data behaves. Data transforms through chunking, embedding, retrieval, and generation. At each step, governance and security become inseparable. The ISM March 2026 update validates this: ISM-2103 requires explicit data owner consent before organisational data processed by AI is used for training or fine-tuning [131]. But ISM-2103 addresses one specific data flow. The broader boundary collapse (classification through chunking and embedding, access control inheritance through RAG pipelines, output classification across security levels) remains unaddressed. For CISOs, data governance maturity is now a direct security input.


Agency data
Classified
Governed
Chunking
Segmented
Classification may not transfer
Embedding
Vectors
Human-readable context lost
RAG retrieval
Mixed sources
Access control inheritance unclear
Generation
New content
Output classification undetermined
Governance dissolves through the pipeline

APS-wide data maturity: 2.02 out of 5. No mandated data governance baseline exists.

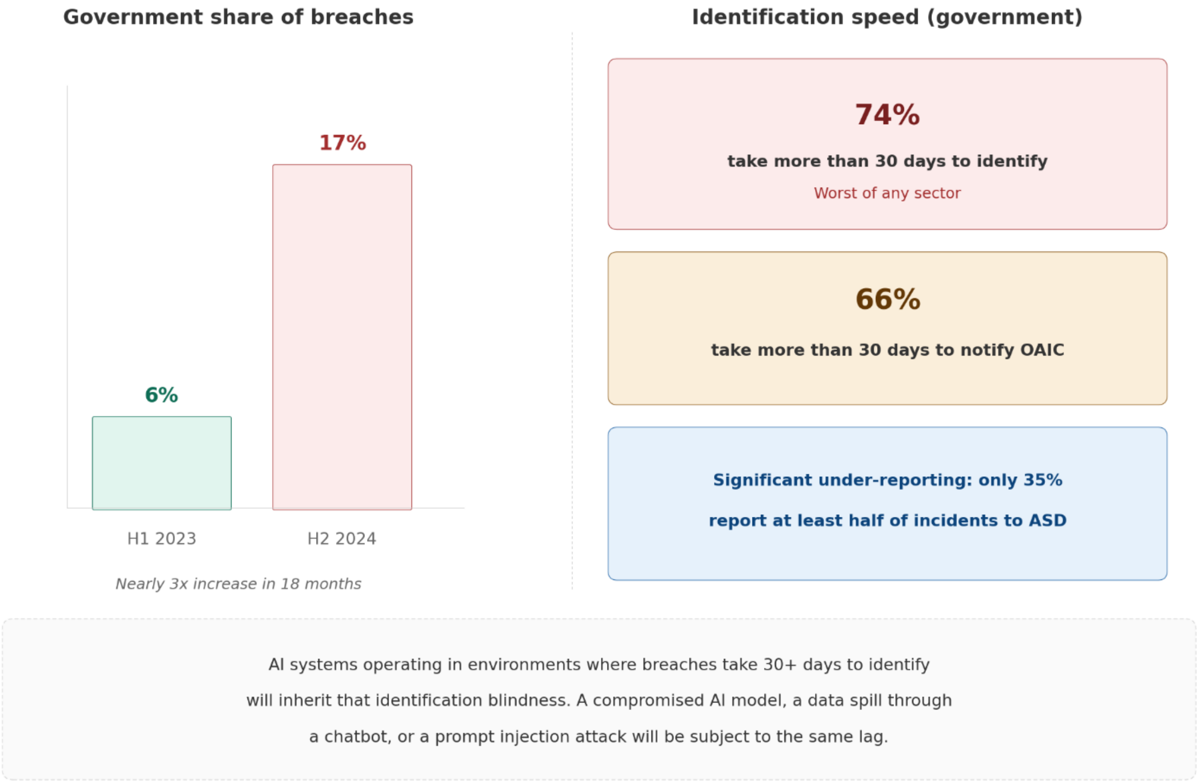
ISM-2103 addresses one data flow. The broader boundary collapse remains unaddressed.

Data governance boundary collapse

5.3 The data breach trajectory

Government data breaches have surged. The Notifiable Data Breaches scheme under the Privacy Act is one reporting channel among several, but it provides the most consistent cross-sector comparison. In the first half of 2023, the Australian Government sector was a small fraction of notifiable breach notifications, outside the top five reporting sectors. By the second half of 2024, it had risen to 100 notifications, representing 17% of all breaches nationally and the second-highest sector [48]. The full year 2024 saw 1,113 total notifications nationally, the highest since the NDB scheme commenced in 2018 [51].

Government agencies are not only experiencing more breaches; they are the slowest sector to detect them. In the second half of 2024, 74% of government breach notifications took more than 30 days to identify, and 66% took more than 30 days to notify the OAIC. Both were the worst figures of any sector [52]. This has not materially improved: the OAIC NDB statistics dashboard records the same 74% identification lag for the second half of 2025 [156]. Government accounts for roughly 46–49% of all incidents ASD responds to nationally [4][5], while being the least capable sector at detecting them. Only 35% of entities reported at least half of their observed incidents to ASD [1], suggesting significant under-reporting on top of slow detection.



AI systems operating in environments where breaches take 30+ days to identify will inherit that identification blindness. A compromised AI model, a data spill through a chatbot, or a prompt injection attack will be subject to the same lag.

Data Breach Surge and Identification Speed

For AI readiness, the detection speed finding is the critical data point. AI systems operating in environments where breaches take more than 30 days to identify will inherit that detection blindness. A compromised AI model, a data spill through a chatbot, or an adversarial prompt injection attack will be subject to the same detection lag.

5.4 The structural problem: why self-assessment is not enough

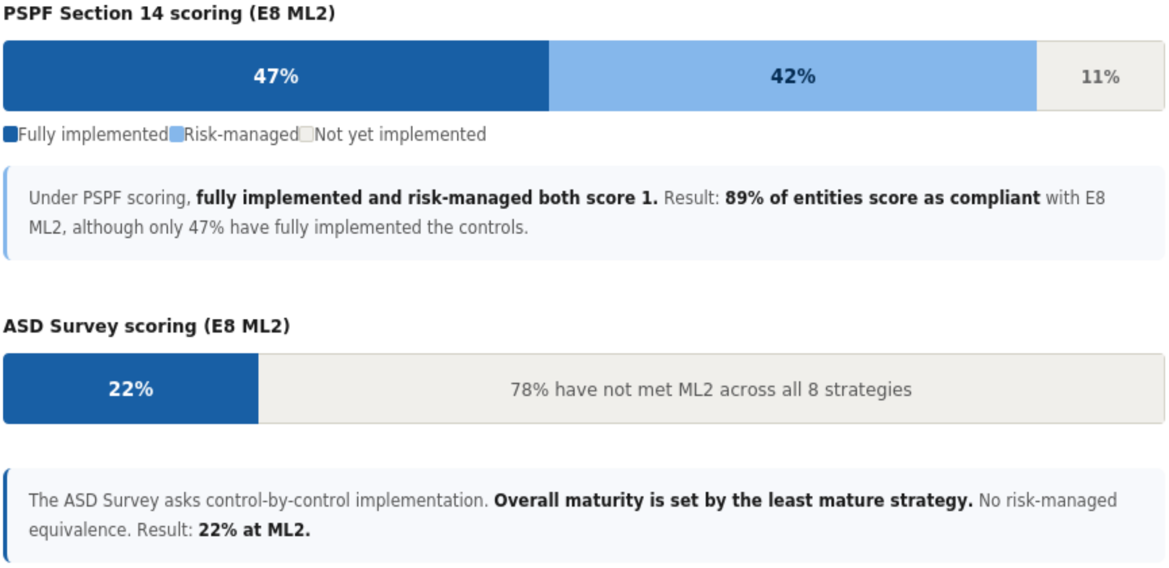
One finding cuts across every audit source and has direct implications for AI governance: self-assessment of security maturity produces structurally optimistic results. The Department of Home Affairs has acknowledged this directly, citing “the optimism bias commonly associated with self-assessment” in the 2024–25 PSPF Assessment Report [42]. In response, PSPF Release 2024 introduced an Assurance Capability designed to support entities to improve their implementation of PSPF requirements and accuracy of reporting. Home Affairs will pilot this capability against the 2024–25 assessment results [42].

The ANAO’s audit record provides the empirical evidence. In 2020–21, ANAO tested entity self-assessments of the Top Four mitigation strategies against actual implementation and found two of three entities that reported full implementation had done so inaccurately [151]. Three years earlier, ANAO had recommended that AGD, ASD and Home Affairs strengthen “the processes for verifying the accuracy of entities’ self-assessment” [150], and identified shortcomings in the Essential Eight Maturity Model that “could lead to entities inadvertently overstating their cyber security compliance” [150]. Across four performance audits from 2013–14, only 4 of 14 non-corporate entities complied with mandatory controls [19]. ANAO’s conclusion: previous audits have not found significant improvement over time, and this pattern continues [18].

The PSPF’s own reporting data illustrates the structural mechanism. Under PSPF Section 14 — Cyber Security Strategies, which requires implementation of the Essential Eight to Maturity Level 2 — only 47% of entities reported full implementation in 2024–25 [42]. A further 42% reported a “Risk-Managed” approach, meaning they have not fully implemented E8 ML2 but have documented mitigations and risk treatment plans. Under the PSPF scoring methodology, both categories score identically: Risk-Managed and Fully Implemented each receive a score of 1 [42]. The result is that 89% of entities score as compliant with the E8 ML2 requirement under the PSPF. The ASD Survey, asking the same agencies about the same requirement on a control-by-control basis where the overall maturity level is set by the least mature strategy, finds 22% at Maturity Level 2 [1]. Both are entity self-reported. The divergence is not between self-assessment and independent verification; it is between two self-assessment instruments with fundamentally different measurement methodologies applied to the same requirement. The PSPF measures whether entities have addressed the requirement through implementation or documented risk management; the ASD Survey measures whether the specific controls are actually in place.

Same requirement, different self-assessment

PSPF Section 14 (Cyber Security Strategies) vs ASD Survey — both entity self-reported, 2024–25



Under PSPF scoring, **fully implemented and risk-managed both score 1.** Result: **89% of entities score as compliant** with E8 ML2, although only 47% have fully implemented the controls.

The ASD Survey asks control-by-control implementation. **Overall maturity is set by the least mature strategy.** No risk-managed equivalence. Result: **22% at ML2.**

Same agencies. Same requirement. Both self-reported. The divergence is methodological:
the PSPF measures whether entities have *addressed* the requirement; the ASD Survey measures whether controls are *in place.*

Sources: PSPF Assessment Report 2024–25, Section 14 [42]; ASD Posture Report 2025 [1]

Same requirement, different measurement methodologies

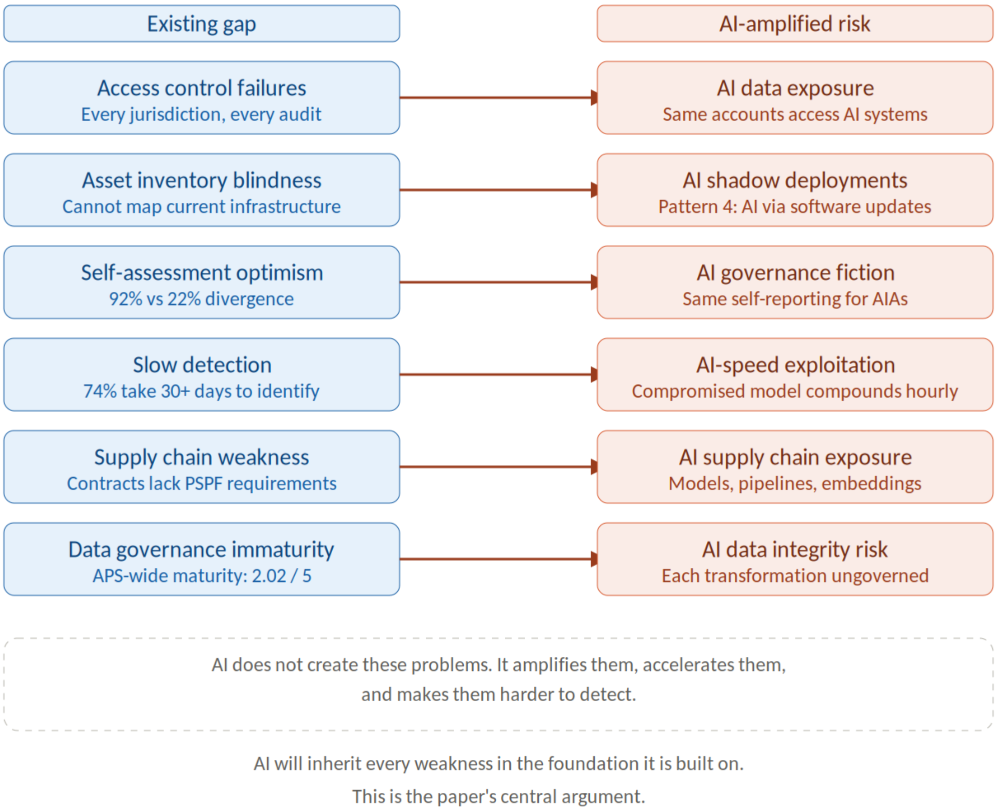
The self-assessment gap is structural, not incidental. Compliance evidence across most Commonwealth entities relies on periodic manual assessments rather than continuous, system-verified data. Assessment boundaries are open to interpretation. ANAO’s audit capacity covers only a handful of agencies per cycle. As early as 2017–18, ANAO recommended building verification processes for entity self-assessments [150]. In 2020–21, ANAO tested those self-assessments and confirmed they overstated compliance [151]. Home Affairs' introduction of the PSPF Assurance Capability in 2024 is the first structural response to a problem documented across seven years of audit findings.

The lesson must transfer to AI governance before the gap opens. AI Impact Assessments, AI Transparency Statements, and Responsible AI Maturity Assessments, as currently designed, all rely on the same self-assessment model. The pattern extends to the instrument literally named "assurance": based on published information, DTA's Pilot AI Assurance Framework is structured self-assessment with no visible indication of an external assessor, independent verification mechanism, or independent review built into the framework [153]. Its security section (Section 6.3, Authority to Operate) defers entirely to PSPF and ISM engagement with the agency's ITSA [153]. If self-assessment produces structurally optimistic results for cyber security — a domain with specific technical controls that can be tested — it will produce the same pattern for AI governance, where the controls are less mature, the assessment criteria are less defined, and the failures are harder to detect. A subtly biased model, a data exposure through a chatbot, or a compromised decision-making pipeline are less visible than conventional breaches and slower to surface.

> **Recommendation 3 (Next):** Mandate AI security posture as a standing Posture assessment category. This creates an independent data point on AI security readiness before self-assessment patterns embed. *(Home Affairs)* [S5]
>
> **Recommendation 4 (Next):** Develop an assurance function for AI Impact Assessments (AIA), incorporating lessons from the PSPF Assurance Capability and documented self-assessment optimism bias [42]. *(DTA)* [S5]

5.5 The amplification effect: what this means for AI adoption

The security reality documented in this section is not a legacy IT problem that sits alongside AI adoption. It is the foundation on which AI is being built. Six causal mechanisms connect today's audit findings to tomorrow's AI security failures:



Existing gap

AI-amplified risk

Access control failures
Every jurisdiction, every audit

AI data exposure
Same accounts access AI systems

Asset inventory blindness
Cannot map current infrastructure

AI shadow deployments
Pattern 4: AI via software updates

Self-assessment optimism
92% vs 22% divergence

AI governance fiction
Same self-reporting for AIAs

Slow detection
74% take 30+ days to identify

AI-speed exploitation
Compromised model compounds hourly

Supply chain weakness
Contracts lack PSPF requirements

AI supply chain exposure
Models, pipelines, embeddings

Data governance immaturity
APS-wide maturity: 2.02 / 5

AI data integrity risk
Each transformation ungoverned

AI does not create these problems. It amplifies them, accelerates them,
and makes them harder to detect.

AI will inherit every weakness in the foundation it is built on.
This is the paper's central argument.

The amplification effect: six pathways from existing gaps to AI-amplified risk

6. Current Cyber Security Policies, Programs & Governance

Australia’s policy architecture has undergone its most significant transformation since 2018. Eighteen months ago, the country had no standalone cyber legislation, no mandatory zero trust requirements, no AI policy in the PSPF, and no formal CISO accountability model. All now exist.

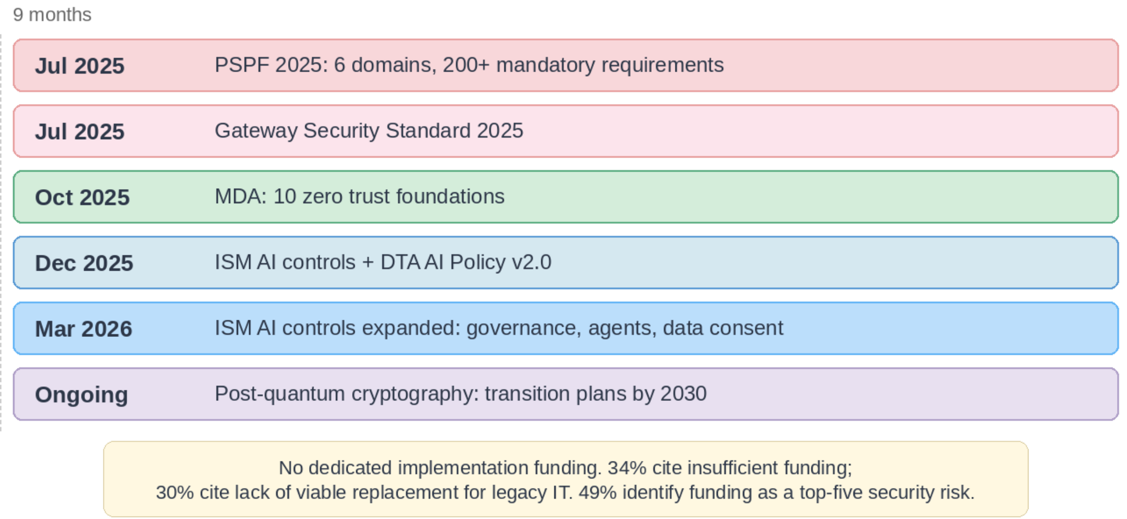
6.1 The policy architecture

Horizon 1 of the 2023–2030 Cyber Security Strategy is substantially complete [30], delivering the Cyber Security Act 2024, a restructured PSPF with over 200 mandatory requirements, zero trust and post-quantum mandates, and MDA [7]. Horizon 2 (2026–2028) is in co-design. This is the window for embedding AI security for both the public sector and broader Australian community.

International experience is instructive. The US built comprehensive AI governance through Executive Order 14110; when it was revoked in January 2025, governance survived only where separately embedded in Office of Management and Budget (OMB) memoranda [87][88][89]. Canada embedded AI requirements into existing Treasury Board directives, mandatory since 2019 and updated multiple times since 2019, the most durable Five Eyes model [84]. The take away is that governance woven into existing architecture survives; parallel structures are fragile.

> **Recommendation 1 (Next):** Integrate Government AI security into the Strategy Horizon 2 co-design. This ensures AI security is embedded in the national cyber strategy as a visible commitment alongside the government's AI adoption ambitions, rather than developed as a separate workstream. *(Home Affairs)* [S6]

6.2 PSPF 2025: progress and strain



9 months

Jul 2025	PSPF 2025: 6 domains, 200+ mandatory requirements
Jul 2025	Gateway Security Standard 2025
Oct 2025	MDA: 10 zero trust foundations
Dec 2025	ISM AI controls + DTA AI Policy v2.0
Mar 2026	ISM AI controls expanded: governance, agents, data consent
Ongoing	Post-quantum cryptography: transition plans by 2030

No dedicated implementation funding. 34% cite insufficient funding;
30% cite lack of viable replacement for legacy IT. 49% identify funding as a top-five security risk.

Reform pile-up

The scale of concurrent reform is visible in the figure above. Within nine months, agencies absorbed the PSPF's expansion to over 200 mandatory requirements [24], two new mandatory standards (Gateway Security Standard 2025 and Systems of Government Significance) [28][29], MDA's ten zero trust foundations [7], two consecutive material ISM updates adding AI-specific controls [131], DTA AI Policy v2.0 [31], and the beginning of post-quantum cryptography transition. Each addresses a genuine gap. Together, they represent a reform burden that agencies documenting 78% non-compliance with the pre-existing baseline are being asked to absorb without (at this stage) dedicated implementation funding.

The mandated CISO role, with specified executive reporting lines, expanded responsibilities, and formalised accountability, strengthens the governance chain. Agency heads remain the risk owners. Self-assessment remains the primary assurance mechanism with documented optimism bias [42], now being addressed through Home Affairs' pilot of the PSPF Assurance Capability [42]. One reform in the pile-up deserves separate attention for AI security. The Systems of Government Significance (SoGS) Standard [29] creates a criticality-based prioritisation of the Commonwealth's most important digital functions and systems, assessed by the potential for significant consequences to economic prosperity, social cohesion, or national interest if disrupted [147]. Declared systems carry additional obligations to mitigate risk and uplift cyber capability. This is existing infrastructure that directly answers the prioritisation question this paper raises throughout: which AI deployments warrant the highest security threshold? AI deployed on or connected to a SoGS-declared system inherits that system's criticality. The SoGS register, still being built through agency assessment, provides a ready-made top tier for AI security prioritisation across Recommendations 2, 8, and 11 without requiring new classification criteria from scratch.

6.3 The AI governance layer: sound design, sequencing risk

The AI governance architecture is taking shape. DTA Policy v2.0 mandates AI Impact Assessments for high-stakes use cases, with the first new mandatory requirements from June 2026 and remaining requirements from December 2026 [31]. Agencies are expected to appoint Chief AI Officers in 2026 [33]. AI Accountable Officials are designated. AI Transparency Statements and agency-level AI registers are being stood up. The ISM in March 2026 now makes it clear that Government agency executive leadership is accountable for ensuring AI systems are secure, controllable, and human-supervised.

The design is sound. The AI Impact Assessment (AIA) is a governance instrument that refers agencies to the ISM for security requirements. ASD has been building AI-specific controls into the ISM through two consecutive material quarterly updates (December 2025 and March 2026), with the detail covered in Section 7. Significant deployment-layer gaps remain in model provenance, prompt injection defence, RAG pipeline security, AI-specific event logging, and AI supply chain assurance.

Canada’s Algorithmic Impact Assessment has been mandatory since 2019, updated multiple times since 2019, the most durable model in the Five Eyes. Australia’s mandatory assessments begin seven years later [84]. The UK’s AI Playbook uses mandatory language linking AI security compliance directly to its existing Government Cyber Security Standard, creating binding requirements without new legislation [82]. Australia could adopt the same approach through the PSPF.

> **Recommendation 2 (Now):** Establish a transitional AI Security Lab to work with agencies moving AI from proof of concept to production, bridging the gap until frameworks, workforce, and assurance mechanisms mature. Initial priority: AI deployments on Systems of Government Significance and GovAI Chat. *(ACSC / DTA)* [S8]

6.4 The funding reality

Australia has committed headline sums: $9.9 billion for REDSPICE, $586.9 million for the 2023–2030 Cyber Security Strategy [46][30]. ASD’s budget has more than doubled, from $1.17 billion in 2021–22 to $2.478 billion in 2025–26, and its workforce has nearly doubled to over 4,000 staff [47][119]. These are real capability investments.

But the investment is heavily concentrated in ASD. Outside ASD, 34% of entities cite insufficient dedicated funding as the most significant reason for continued reliance on legacy IT that impedes Essential Eight implementation, and a further 30% cite lack of a viable replacement [1]. Under the PSPF Assessment, 49% of entities identified funding, resources, or capability limitations as a top-five security risk [42]. The counter-argument is that 91% of entities have planned improvement programs and 83% of those have funded them [1]. This sounds reassuring until tested against results: if 91% have plans and 83% have funding, yet 78% cannot meet the pre-existing E8 baseline after three years, the funded programs are evidently not adequate to the task. Having a funded program is not the same as having adequate funding.

It also shows the assurance limitations of having activity-based reporting (do agencies have planned and funded improvement programs) versus verified outcome-based reporting (are those programs funded and commensurate with reducing risk), as outlined in Section 7.2.

The PSPF 2025 reforms mandate zero trust, AI governance, and post-quantum preparation with no visible dedicated implementation funding [24]. The 2025–26 Budget contained no significant new standalone cyber measures [118]. KPMG's budget analysis noted the absence of new investment in important emerging technology fields, including AI [117]. This was the same year the PSPF expanded to over 200 mandatory requirements.

The funding conversation also needs to distinguish between three lanes. First, IT modernisation: replacing legacy systems, moving to cloud, adopting modern architectures. This is a productivity investment that delivers security improvement as a byproduct. Agencies that modernise their IT estate automatically improve their security posture because modern platforms come with stronger default controls. Second, dedicated cyber investment: the people, tools, and processes to secure whatever IT estate the agency operates. Third, structural model: whether over 190 entities each independently funding and managing cyber security capabilities is sustainable, or whether shared capability models (as the government has accepted for AI through GovAI Chat) would deliver better outcomes. The Cyber Hubs concept was designed to address this third lane. It was not replaced.

Individual agency uplift programs run to substantial sums. Services Australia's $1.8 billion modernisation program includes cyber uplift; DFAT received $227.8 million for cyber resilience [45]. But these are agency-specific, funded within individual appropriations. No single published aggregate of total government cyber spending outside ASD exists. The funding model is fragmented and structurally misaligned with a threat environment that does not respect organisational boundaries.

> The Cyber Hubs program demonstrated the problem. Piloted with four Hub Lead Agencies from April 2021, it identified 42 core services that could be delivered as shared capabilities to smaller agencies. An ANAO performance audit identified legitimate procurement and operational concerns, funding was not extended, and the 2023–2030 Cyber Security Strategy contains no mention of Cyber Hubs [30]. Cyber security responsibility reverted to individual agency budgets. The structural problem Cyber Hubs was designed to solve has not gone away. Over 190 entities each independently funding cyber security capabilities is inherently duplicative, and the smaller agencies that Hubs were designed to serve are precisely those that most need shared capability.
>
> This matters because the government has, in parallel, accepted the shared-service logic for AI that it abandoned for cyber security. GovAI Chat is a $225.2 million whole-of-government platform designed precisely so that smaller agencies that may lack the scale and resourcing to adopt native AI tools can access AI capability through a shared service [33][34]. The same economies-of-scale argument applies to cyber security, and cyber security is a prerequisite for safe AI deployment. The government is building a shared AI capability on a foundation where shared cyber capability was attempted and not replaced.

> **Recommendation 5 (Next):** Make cyber security uplift funding ongoing at both agency and whole-of-government level, rather than treating it as one-off programs. AI security requirements should be built into this funding as they emerge, and AI security budget lines should be required in all new AI deployment business cases above a defined threshold. Modernisation investment that replaces legacy IT delivers security improvement as a byproduct; dedicated cyber funding addresses the residual gap. Both are prerequisites for secure AI adoption. *(Finance / Home Affairs / ACSC)* [S6]

Assessment

The policy architecture can support AI security. Five Eyes coordination is producing operationally useful guidance [13][14]. And for the first time, the structural conditions for AI security governance are in place. The ISM March 2026 update establishes executive accountability for AI security (GOV-08). The PSPF 2025 reforms mandate the CISO role with direct executive reporting lines and expanded responsibilities. These are not incremental changes. Together, they create an accountability chain from technical controls through to executive risk ownership that did not exist eighteen months ago.

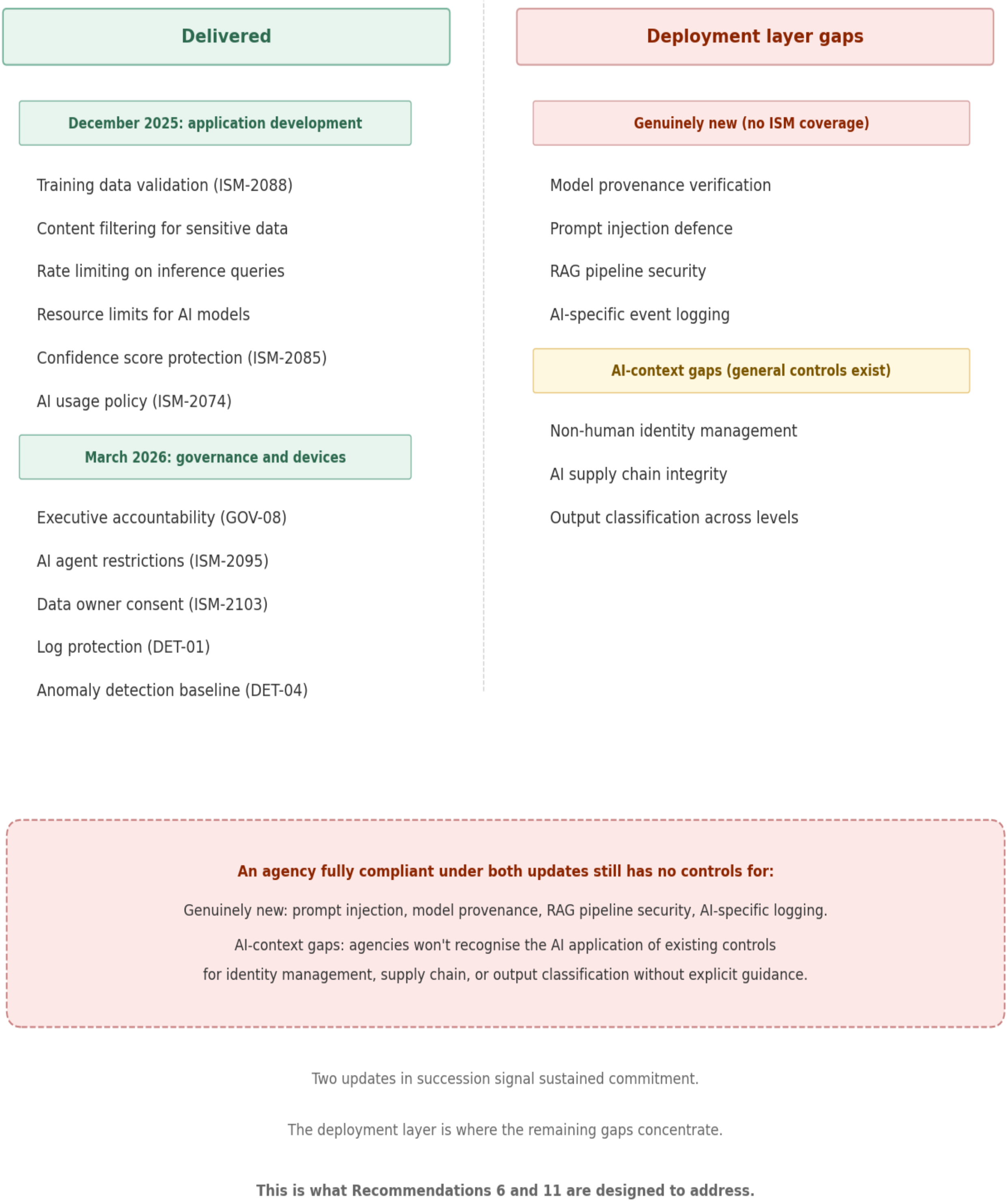
For CISOs, this is the moment to bring AI security and traditional ICT security reporting together, addressing both the parallel concerns and the amplification risks this paper documents, so that executives can make informed decisions about AI deployment risk appetite with a complete picture. The policy settings are sound. The capacity to implement them remains the defining gap: modernisation of the legacy IT that impedes security improvement, dedicated investment in the security workforce and tools, and a structural model that does not require 190+ agencies to independently build the same capabilities.

7. Current Cyber Security Frameworks & Standards

Australia’s technical security backbone rests on the ISM as the control library, the Essential Eight as the current compliance baseline, and MDA as the architectural direction.

7.1 The ISM: a risk-based control library, not a compliance checklist

The ISM is not a fixed checklist: agencies select controls based on their risk context, data classification, and systems [6]. It is updated quarterly, and ASD has used that quarterly cadence to begin embedding AI security into the ISM through two consecutive material updates.



Delivered

December 2025: application development

- Training data validation (ISM-2088)
- Content filtering for sensitive data
- Rate limiting on inference queries
- Resource limits for AI models
- Confidence score protection (ISM-2085)
- AI usage policy (ISM-2074)

March 2026: governance and devices

- Executive accountability (GOV-08)
- AI agent restrictions (ISM-2095)
- Data owner consent (ISM-2103)
- Log protection (DET-01)
- Anomaly detection baseline (DET-04)

Deployment layer gaps

Genuinely new (no ISM coverage)

- Model provenance verification
- Prompt injection defence
- RAG pipeline security
- AI-specific event logging

AI-context gaps (general controls exist)

- Non-human identity management
- AI supply chain integrity
- Output classification across levels

An agency fully compliant under both updates still has no controls for:

Genuinely new: prompt injection, model provenance, RAG pipeline security, AI-specific logging.

AI-context gaps: agencies won't recognise the AI application of existing controls for identity management, supply chain, or output classification without explicit guidance.

Two updates in succession signal sustained commitment.

The deployment layer is where the remaining gaps concentrate.

This is what Recommendations 6 and 11 are designed to address.

ISM AI controls

December 2025 introduced the first AI-specific controls in an entirely new AI application development section. March 2026 expanded coverage from application development into governance and device-level restrictions [131]. The detail is in the figure above; the analytical point is what follows.

This progressive approach is the right one. Two updates in succession signal sustained ASD commitment, not a one-off response. The consultation-led approach is how good technical controls get built. It is also why it takes time. Each quarterly update cycle involves drafting, stakeholder consultation, and publication; agencies then need to interpret, budget for, and implement the new controls across diverse environments. For a framework that needs to cover everything from a small agency running a summarisation tool on low-sensitivity data to a large department deploying AI-assisted decision-making on sensitive citizen records, getting the controls right matters more than getting them fast. The coverage, however, has limits, and AI adoption is not waiting for the ISM to complete its build. What remains is a mix of genuinely new gaps and AI-context gaps. Some deployment-layer risks have no ISM coverage at all: model provenance verification, prompt injection defence, RAG pipeline security, and AI-specific event logging are not addressed by any existing control, even conceptually. Others, such as non-human identity management and AI supply chain integrity, fall within the scope of existing general controls, but agencies may not recognise the application without explicit ASD guidance. The outcome-based approach recommended in this paper (Rec 6) would address both: new controls for genuinely new risks, and explicit AI context for controls that already exist in principle. An agency fully ISM-compliant under both updates has an AI usage policy, protects training data, filters outputs, and has executive accountability, but has no specific ISM controls covering what happens when an approved AI system retrieves, processes, and acts on sensitive data across agency boundaries in production.

ASD has published eight AI security publications across 2024–2026 covering development, deployment, data security, supply chain, OT integration, general engagement, and small business guidance [11], and co-authored two major Five Eyes joint publications [13][14]. The content is substantial. The gap is in format and integration: these are advisory guidance, not numbered assessable controls integrated into ISM control families, connected to IRAP scope, or referenced as mandatory under the PSPF. For advisory guidance, this is a translation and integration problem, not a content creation problem. For genuinely new deployment-layer risks, the controls need to be built.

As AI security controls continue to be embedded in the ISM, the approach should define security outcomes (model provenance must be traceable, training data integrity must be verifiable, AI outputs must be classifiable) rather than prescribe specific technical implementations. AI’s rate of technical change means prescriptive controls will lag the technology between quarterly ISM updates. Outcome-based controls provide stability; ASD’s advisory guidance suite can address evolving technical implementation without waiting for formal ISM cycles. International precedent validates this: NIST’s Cybersecurity of AI Systems (COSAiS) project is taking the same approach for SP 800-53 [139], and MITRE’s Securing AI Frameworks and Environments (SAFE-AI) report identified 100 SP 800-53 controls as “potentially AI-affected” [140].

> **Recommendation 6 (Next):** Continue embedding AI security controls into the ISM, extending coverage from application development and governance into the operational deployment layer: model provenance, prompt injection defence, AI-specific logging, RAG pipeline security, agent protocol security, and the data governance-security intersection documented in Section 5.2, as outcome-oriented

> controls. Technical implementation guidance should sit alongside the ISM and update at the pace the technology demands. *(ACSC)* [S7]

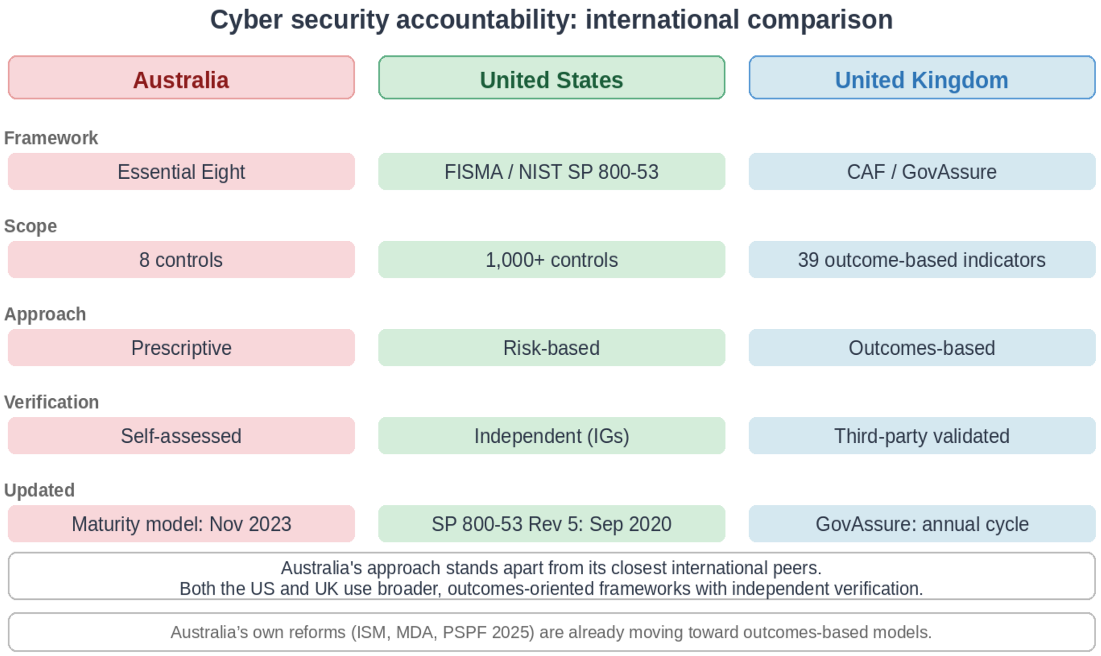
7.2 The Essential Eight: evolving the accountability model

The Essential Eight has done what it was designed to do. It translated complex security requirements into a form executives could understand, auditors could measure, and agencies could track. It gave the Commonwealth its first mandatory cyber security baseline. The compliance data in Section 5 confirms that most agencies have not yet achieved it, and until they do, the E8 remains the foundation.

But the E8 is now embedded well beyond its original purpose. It anchors ANAO audit programs, the annual Posture Report, PSPF Policy 10 compliance (now the Technology domain), and parliamentary reporting cycles. That institutional embedding is both its strength (it created visibility and accountability where none existed) and the reason evolution is complex. Any change to the accountability mechanism flows through audit methodology, reporting frameworks, and agency planning cycles simultaneously.

ASD’s own actions signal the direction of travel. The four heightened areas of focus for board-level governance in 2025–26, published jointly with the AICD, are event logging, legacy IT management, supply chain risk, and post-quantum cryptography preparation [144]. All four extend beyond what E8 measures at the mandatory ML2 baseline. MDA’s ten foundations address identity, monitoring, and secure design that E8 was never designed to cover. New AI controls are being built into the ISM, not the Essential Eight. The E8 Maturity Model was last substantively updated in November 2023 [8]; its parent publication has not been updated since 2017 [9]. The framework that matters most for accountability is receiving the least investment in evolution.

The international comparison reinforces the point.



Cyber security accountability: international comparison

	Australia	United States	United Kingdom
Framework	Essential Eight	FISMA / NIST SP 800-53	CAF / GovAssure
Scope	8 controls	1,000+ controls	39 outcome-based indicators
Approach	Prescriptive	Risk-based	Outcomes-based
Verification	Self-assessed	Independent (IGs)	Third-party validated
Updated	Maturity model: Nov 2023	SP 800-53 Rev 5: Sep 2020	GovAssure: annual cycle

Australia's approach stands apart from its closest international peers.
Both the US and UK use broader, outcomes-oriented frameworks with independent verification.

Australia’s own reforms (ISM, MDA, PSPF 2025) are already moving toward outcomes-based models.

International cyber accountability comparison

Australia’s approach stands apart from its closest international peers. Both the US and UK use broader, outcomes-oriented frameworks with independent verification: FISMA and NIST SP 800-53 in the US [145], and the Cyber Assessment Framework (CAF) operationalised through GovAssure in the UK [146]. This is not an argument to replace E8 overnight. It is an observation that Australia’s own reforms (ISM risk-based controls, MDA outcomes-based architecture, PSPF 2025 expanded scope) are already moving in the same direction. The accountability mechanism has not kept pace with the reforms it is supposed to measure.

In practice, the evolution is already underway. The ISM, MDA, and PSPF 2025 are all building beyond the eight-control model. The Essential Eight remains as eight technical controls within the ISM; what has not yet evolved is the accountability mechanism that measures and reports on the broader security posture. The direction of travel is clear: from narrow prescriptive controls toward a risk-based, outcome-oriented accountability model, as the US and UK have already adopted.

> **Recommendation 10 (Next):** Evolve the Essential Eight assessment into a broader cyber security accountability mechanism that accommodates the reforms already delivered: ISM AI controls, MDA foundations, PSPF 2025 expanded requirements, and AI security indicators. The current model measures eight controls through a prescriptive maturity lens; the reforms it needs to capture are risk-based and outcome-oriented. This should include extending Posture assessment to measure outcomes (can the agency detect, trace, and respond?) rather than inputs alone (does the agency have a plan?). The Essential Eight would remain as the foundational technical baseline within the broader mechanism; agencies that have not yet achieved it still need to. The evolution is in what gets measured and reported alongside it, through a consultation and transition process that allows ANAO audit methodology, Posture reporting, and agency planning cycles to adapt. *(ACSC / Home Affairs)* [S7]

7.3 MDA, zero trust, and AI

Modern Defensible Architecture (MDA) introduced ten technology-agnostic foundations combining zero trust with secure-by-design [7]. AI workloads are precisely the use case zero trust was designed for. Consider what a typical AI deployment requires: broad data access across agency holdings for RAG queries, API integrations to external model providers, non-human identities (AI service accounts) operating at machine speed, and supply chain dependencies on model providers whose security posture the agency does not control. Every one of these characteristics violates the implicit trust assumptions that conventional network architectures rely on. Zero trust’s core principles (verify explicitly, use least privilege, assume breach) map directly to AI deployment security requirements.

What does not yet exist is MDA implementation guidance specific to AI deployment architectures: identity verification for AI service accounts and non-human identities, micro-segmentation of AI data pipelines to enforce classification boundaries, continuous verification of AI model integrity to detect supply chain compromise or model drift, and access control architectures that distinguish between what data a model can access for training versus inference versus fine-tuning.

> **Recommendation 9 (Next):** Extend MDA to AI workloads. This ensures AI systems operate under the same zero trust principles the government has mandated for

> conventional systems, rather than inheriting the implicit trust assumptions those systems are moving away from. *(ACSC)* [S7]

7.4 IRAP and the assurance pathway

Australia's security assurance model for government technology runs through IRAP. As AI systems move into production handling sensitive government data, they need to fall within IRAP's assessment scope. Today, IRAP does not explicitly cover AI-specific risks: model provenance, training data integrity, prompt injection resilience, or AI supply chain assurance. Even where AI risks could reasonably fall within an assessment's scope, the realities of budget constraints, customer-defined scope boundaries, and assessor knowledge mean they may not be examined in practice. An AI system can pass an IRAP assessment of its hosting infrastructure while the AI-specific risk profile remains unexamined.

The path for Australian government is evolving IRAP to incorporate AI-specific assessment criteria for the platform and AI workload level, not layering a new mandatory certification on top. This is distinct from the deployment-layer assurance gap discussed below: IRAP addresses whether the platform and its AI services meet security requirements at a point in time; Recommendation 11 addresses the continuous assurance needed for the deployment layer where AI meets agency data in production.

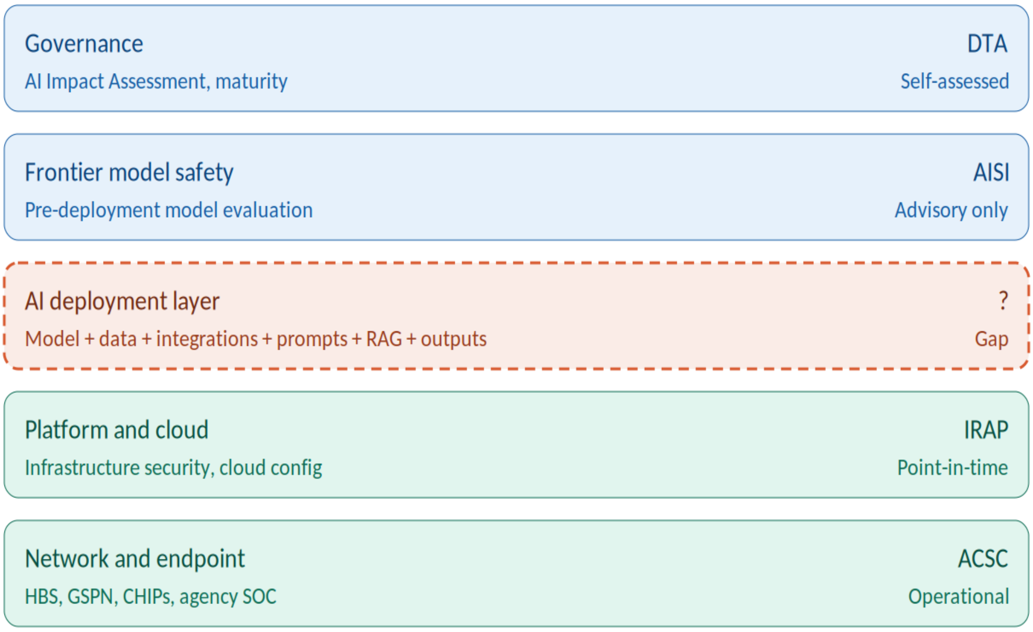
> **Recommendation 7 (Next):** Evolve IRAP to cover AI workloads, expanding assessment scope to include AI-specific risks at higher sensitivity levels. The IRAP program review underway is the vehicle. *(ACSC)* [S7]

7.5 The authority to operate gap

IRAP is an assessment mechanism. The authority to operate (ATO) is a decision mechanism. In Australia, ATO decisions are agency-led and typically point-in-time. AI breaks this model. Models update on weeks-to-months cadence, RAG pipelines evolve, new attack techniques emerge dynamically. A point-in-time ATO for an AI deployment is an assessment of a system that no longer exists.

Anthropic's Anthropic-designated GTG-1002 disclosure [142] illustrated how both layers fail together. Actors assessed by Anthropic with high confidence as Chinese state-sponsored used task decomposition and false context framing to manipulate Claude Code into conducting what the model interpreted as legitimate defensive security research. This was a prompt-level bypass, not an infrastructure exploit. The threat actor targeted roughly thirty global organisations; the AI executed 80–90% of the campaign's tactical operations with minimal human intervention, and succeeded in compromising a small number of targets. But it was the deployment context that determined the blast radius: Claude Code's agentic capabilities (shell execution, file system access, API calls) combined with inadequate access controls, insufficient monitoring, and weak output handling meant the model was manipulated and the deployment layer determined what damage that manipulation could do. Neither a one-off infrastructure assessment nor a pre-deployment model evaluation alone would have detected this. It required continuous assurance across both layers.

Australia's assurance architecture currently leaves the deployment layer unaddressed:


Governance
DTA
AI Impact Assessment, maturity
Self-assessed
Frontier model safety
AISI
Pre-deployment model evaluation
Advisory only
AI deployment layer
?
Model + data + integrations + prompts + RAG + outputs
Gap
Platform and cloud
IRAP
Infrastructure security, cloud config
Point-in-time
Network and endpoint
ACSC
HBS, GSPN, CHIPs, agency SOC
Operational

Neither infrastructure assessment nor model evaluation covers
what happens when AI meets agency data in production.

Applies to both citizen-facing AI and internal platforms like GovAI Chat.

Assurance gap stack

A consistent pattern reinforces the gap across DTA's three principal AI governance instruments: the AI Technical Standard [152], the Pilot AI Assurance Framework [153], and the AI Impact Assessment Tool [154]. All three handle security identically, deferring to ISM and PSPF. None contains AI-specific security criteria for the deployment layer where model, agency data, integrations, prompts, RAG pipelines, and outputs converge. The Technical Standard makes the hierarchy explicit: adversarial testing and Security Authority to Operate are classified as Recommended, while bias, safety, reliability, and compliance testing are all Required [152].

The deployment-layer gap applies to both citizen-facing AI systems (where the attack surface includes adversarial external input from any user) and internal platforms like GovAI Chat, where the risk shifts to insider threats, privileged access inheritance, and cross-agency boundary enforcement. The threat profile differs; the absence of an assessment owner is common to both. For citizen-facing systems, the priority is hardening against prompt injection, data exfiltration, and adversarial manipulation. For internal systems, it is ensuring the AI layer does not inherit the privileged access failures and data governance weaknesses documented in Section 5, and that cross-agency boundaries on shared platforms enforce need-to-know at the AI layer.

No government has mandated continuous AI-specific authorisation. The US is closest: the Federal Risk and Authorization Management Program (FedRAMP) treats AI services identically to any other cloud service, while the US DoD has developed a layered authorisation concept where infrastructure, platforms, and containers each carry their own authorisation, with the AI deployment layer inheriting controls from below and receiving a narrower assessment on top. Continuous ATO is operational for software delivery through DoD Platform One (over 700 active pipelines) but has never been applied to AI workloads. The FY2026 NDAA (Sections 1512–1513) mandates AI-specific cybersecurity frameworks for the Department of Defense by mid-2026, and NIST COSAiS is developing AI control extensions for SP 800-53. These will converge, but they have not converged yet.

Singapore's GovTech offers the most advanced operational tooling through its AI Guardian platform, integrating pre-deployment testing into CI/CD pipelines with continuous runtime guardrails. Approximately one-third of Singapore government agencies use it daily [157].

The pressure toward continuous authorisation predates AI. Cloud adoption and DevSecOps pipelines were already incompatible with 6–12 month assessment cycles. AI does not merely add to this pressure. It makes continuous authorisation essential rather than aspirational. An agentic AI system making autonomous decisions with access to sensitive government data cannot operate under a three-year-old authority to operate that assessed an infrastructure configuration, not the AI deployment running on top of it. In practice, this means a layered cadence: periodic assessment (annual or longer) for infrastructure baselines and architectural compliance; frequent review (quarterly) for configuration drift, model updates, and access control changes; and continuous automated monitoring of prompt patterns, model behaviour anomalies, data access across classification boundaries, and agent action chains.

> **Recommendation 11 (Next):** Evolve ATO from point-in-time to continuous, prioritising AI workloads on Systems of Government Significance. Separate infrastructure authorisation (periodic) from deployment-layer authorisation (continuous). *(ACSC / Home Affairs)* [S7]

7.6 Risk-tiered AI classification: not all AI carries the same risk

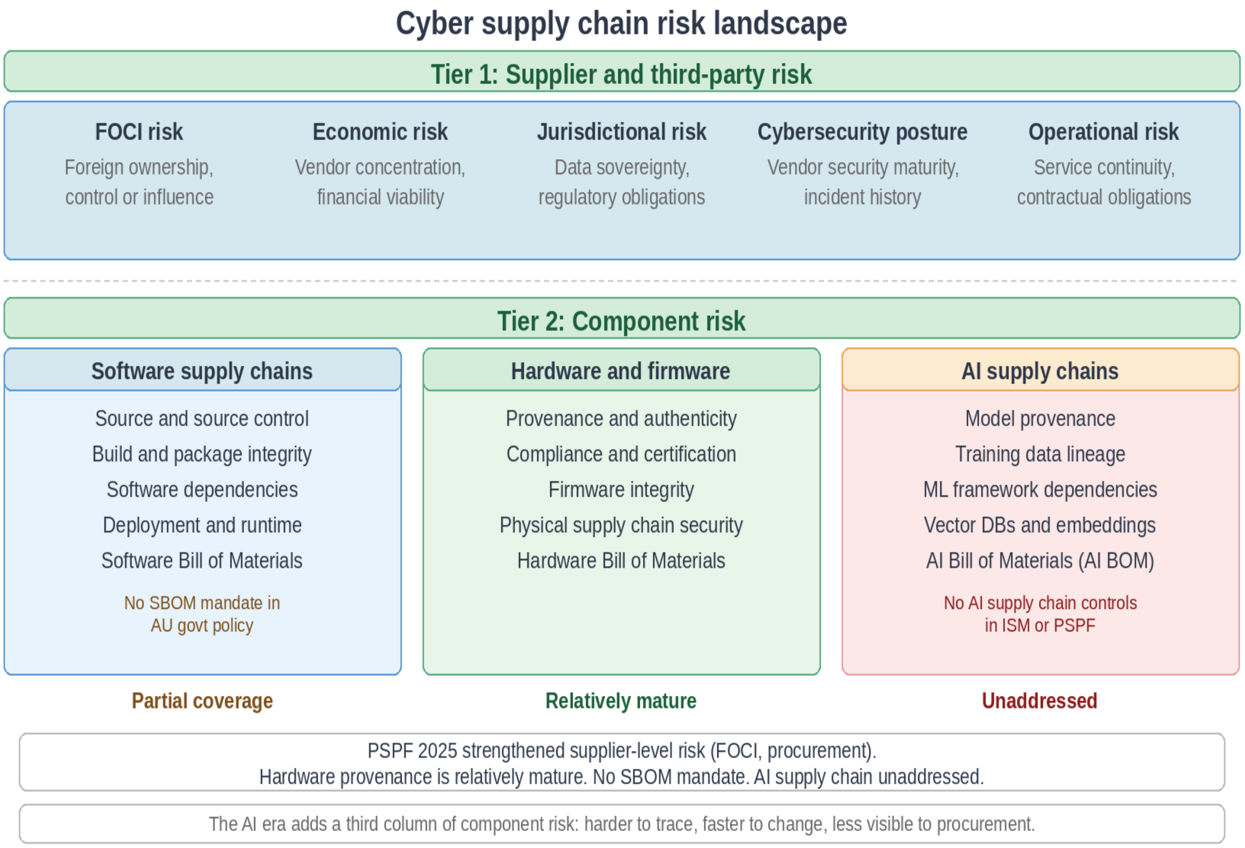
An agency running a summarisation tool on OFFICIAL-level meeting notes presents a fundamentally different risk profile from one deploying AI-assisted decision-making on PROTECTED citizen immigration records. Without a classification model mapping controls to risk levels, agencies face a binary choice: maximum controls everywhere (slowing adoption) or minimum controls everywhere (creating unacceptable exposure).

Tested models exist. AWS's Generative AI Security Scoping Matrix classifies AI workloads across five scopes with mapped controls [99]. Canada's Algorithmic Impact Assessment assigns systems to four impact levels, has been operational since 2019, and updated multiple times since [84]. Combined with Australia's existing data classification framework (OFFICIAL through TOP SECRET), these produce a practical matrix. The DTA's AI Impact Assessment provides the governance trigger; a risk-tiered classification provides the security response.

> **Recommendation 8 (Next):** Adopt a risk-tiered AI classification model that combines AI use-case risk with data classification levels and system criticality. AI deployed on Systems of Government Significance should attract the highest control tier by default. *(ACSC / DTA)* [S7]

7.7 The supply chain gap

Cyber supply chain risk operates at two levels. Supplier risk addresses who the agency is in business with: foreign ownership, economic concentration, jurisdictional exposure, and the vendor's security posture. Component risk addresses what is inside the products: software dependencies, hardware provenance, and for AI systems, model provenance, training data lineage, and ML framework integrity.


Cyber supply chain risk landscape
Tier 1: Supplier and third-party risk
FOCI risk
Foreign ownership, control or influence
Economic risk
Vendor concentration, financial viability
Jurisdictional risk
Data sovereignty, regulatory obligations
Cybersecurity posture
Vendor security maturity, incident history
Operational risk
Service continuity, contractual obligations
Tier 2: Component risk
Software supply chains
Source and source control
Build and package integrity
Software dependencies
Deployment and runtime
Software Bill of Materials
No SBOM mandate in AU govt policy
Hardware and firmware
Provenance and authenticity
Compliance and certification
Firmware integrity
Physical supply chain security
Hardware Bill of Materials
AI supply chains
Model provenance
Training data lineage
ML framework dependencies
Vector DBs and embeddings
AI Bill of Materials (AI BOM)
No AI supply chain controls in ISM or PSPF
Partial coverage
Relatively mature
Unaddressed
PSPF 2025 strengthened supplier-level risk (FOCI, procurement).
Hardware provenance is relatively mature. No SBOM mandate. AI supply chain unaddressed.
The AI era adds a third column of component risk: harder to trace, faster to change, less visible to procurement.

Supply chain risk landscape

The PSPF 2025 reforms strengthened the supplier risk layer. FOCI reporting requirements were expanded [24], and existing procurement frameworks address vendor-level due diligence. Hardware provenance is relatively mature, with government processes for tracing physical supply chains and managing foreign ownership risk.

The component layer is where the gaps concentrate. The ISM includes SBOM requirements for software development (ISM-1730), but no mandatory requirement exists for agencies to demand SBOMs from vendors or model providers through procurement. Software supply chain management is not as advanced as how government addresses hardware provenance and foreign ownership. For AI systems, this gap compounds dramatically: model provenance, training data lineage, ML framework dependencies, vector databases, and embedding models all introduce supply chain risk that traditional procurement processes were not designed to trace. The threat evidence (malicious models on Hugging Face [107], namespace hijacking on major platforms [108]) demonstrates these risks are already being exploited.

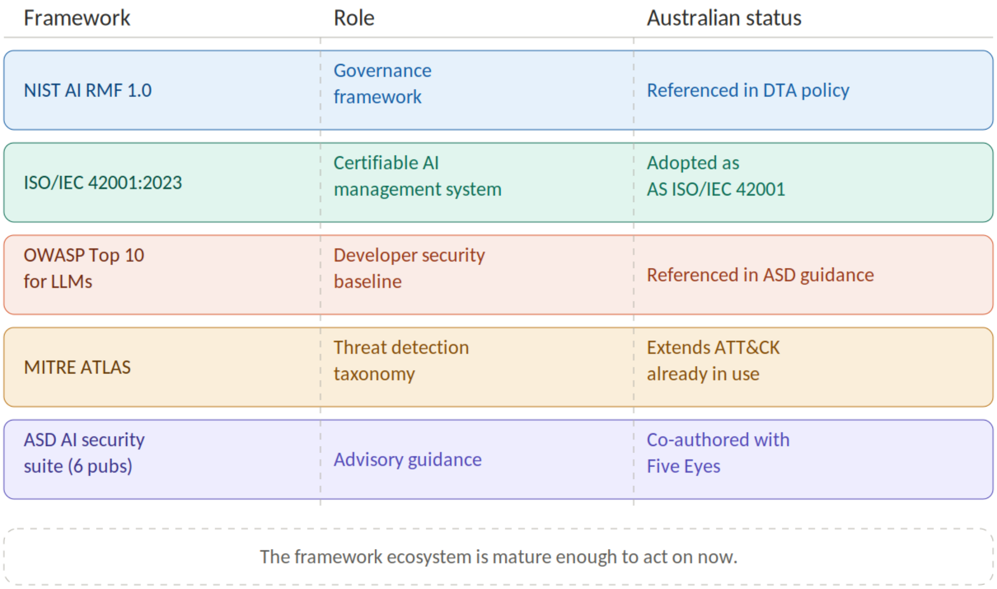
The AI era does not replace the existing supply chain challenge. It adds a third column of component risk alongside software and hardware, with elements that are harder to trace, faster to change, and less visible to conventional procurement. The ISM AI controls (Rec 6) should include model provenance controls, IRAP assessments (Rec 7) should include AI supply chain assurance, and the risk-tiered classification (Rec 8) should escalate supply chain controls with data sensitivity.

> **Recommendation 12 (Next):** Establish AI supply chain transparency requirements: centralised assessment of common foundation models by ACSC, with agencies maintaining deployment-specific model provenance and AI bills of materials (BOM). *(ACSC / Home Affairs)* [S7]

7.8 The international framework ecosystem

Five mature frameworks (NIST AI RMF, ISO 42001, OWASP Top 10 for LLMs, MITRE ATLAS, and ASD's own guidance suite) form a coherent ecosystem that Australian government has already formally referenced. One caveat: the NIST AI RMF is under revision following a July 2025 US White House directive [143]; the practical recommendation is to use NIST for conceptual governance and progress to ISO 42001 for auditable certification.

Approximately 70% of ATLAS mitigations map to existing security controls [115]. This is not a greenfield exercise. But extending controls that 78% of agencies have not yet implemented produces paper coverage, not operational security.



Framework	Role	Australian status
NIST AI RMF 1.0	Governance framework	Referenced in DTA policy
ISO/IEC 42001:2023	Certifiable AI management system	Adopted as AS ISO/IEC 42001
OWASP Top 10 for LLMs	Developer security baseline	Referenced in ASD guidance
MITRE ATLAS	Threat detection taxonomy	Extends ATT&CK already in use
ASD AI security suite (6 pubs)	Advisory guidance	Co-authored with Five Eyes

The framework ecosystem is mature enough to act on now.

Framework Decision Matrix

Assessment

The framework infrastructure can support AI security. The ISM's risk-based architecture supports the progressive embedding of AI controls, and with the December 2025 and March 2026 updates, ASD has begun building AI content through its existing quarterly cadence. MDA addresses AI architectural challenges. IRAP provides an expandable assessment pathway. International frameworks offer tested models. What is needed is translation, integration, and the base working in practice. The 78% of agencies that have not met the pre-AI baseline need modernisation of the legacy IT that constrains their security posture, dedicated investment in the people and tools to operate modern security controls, and efficient models for shared capability. As the accountability mechanism evolves from the Essential Eight toward a broader framework that captures these reforms, that foundation must come first.

8. Current Cyber Security Technology Programs

8.1 What ASD already operates

ASD operates a number of programs forming a multi-layered view of the Commonwealth's cyber threat environment. Host Based Sensor (HBS) deploys endpoint telemetry across participating government entities. Government Security Posture Network (GSPN) provides network-level visibility. Cyber Hygiene Improvement Programs (CHIPs) conducts external vulnerability scanning across government domains, with HOT CHIPs providing rapid-response scanning [1]. Australian Protective Domain Name Service (AUPDNS) provides DNS-level blocking. Cyber Threat Intelligence Sharing (CTIS) provides machine-speed bidirectional sharing, now mandatory under PSPF Direction 003-2024 [27].

These programs provide supplementary national-level visibility. They do not replace agency-level security operations. All were designed for conventional threats.

8.2 The AI security gap and the lab

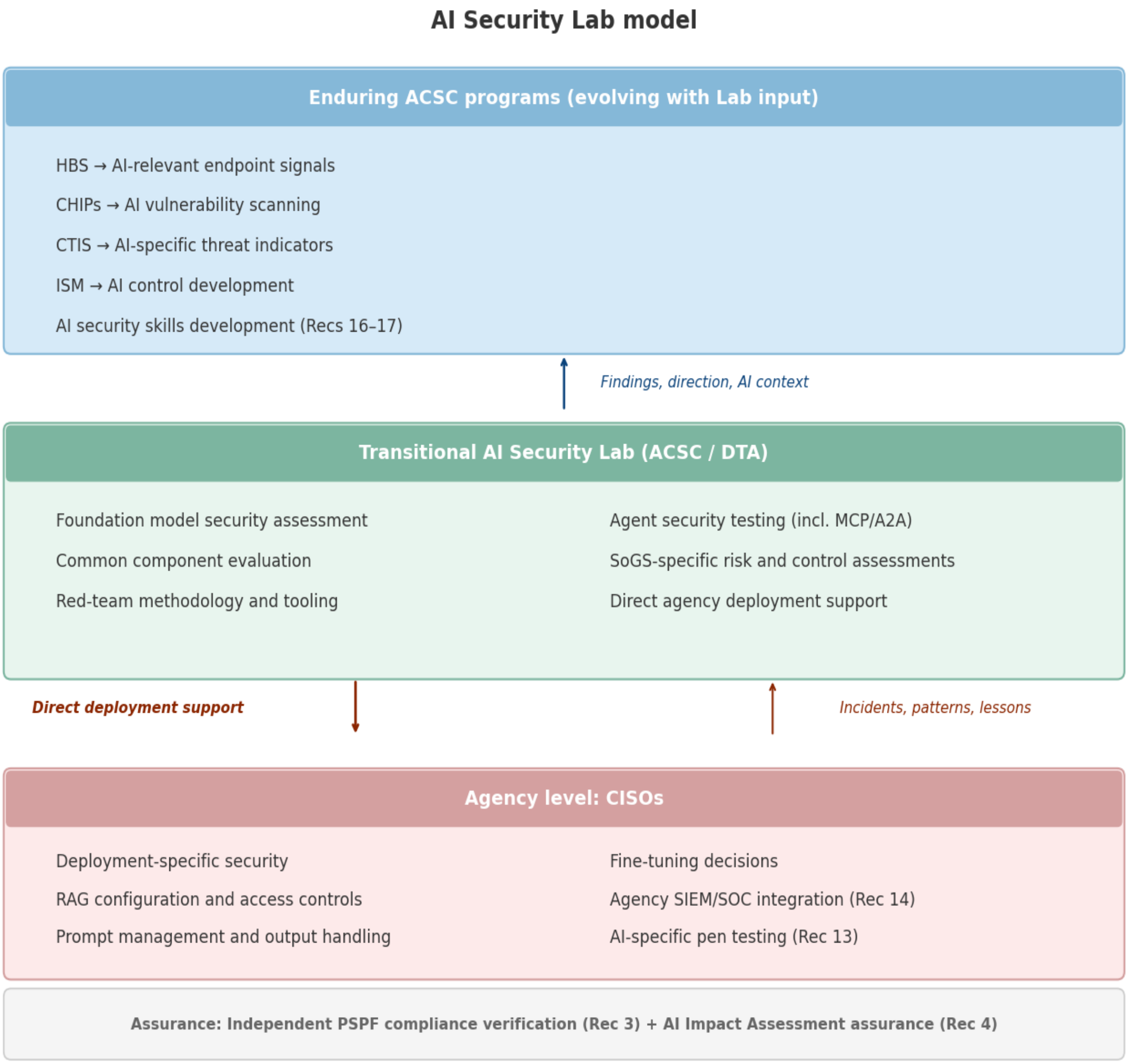
These programs were not designed for AI. The $19 billion value ambition requires agencies to move from proof of concept to production safely. Six structural factors converge: the deployment layer has no assessment owner; AI is entering environments with weak security and data governance; AI security expertise is scarce, and centralising it is more efficient than 190 agencies competing for the same people; and operational findings from a central capability feed framework development. The fifth factor is pace. ASD is engaging with stakeholders and building AI controls into the ISM through a consultation-led quarterly process. That is the right approach for durable technical controls. But the ISM's quarterly cadence, followed by agency interpretation and implementation cycles, operates on a fundamentally different timeline from AI adoption. GovAI Chat trials start in months; agencies are already running AI in production environments. A transitional capability bridges the gap between AI deployment speed and framework development speed. The sixth factor is that budget cycles compound the delay: even when controls are published, agencies need funding to implement them, and insufficient dedicated funding remains among the most significant reasons agencies cannot move off the legacy IT that blocks compliance with the pre-AI baseline.

> **Recommendation 2 (Now):** Establish a transitional AI Security Lab, a blended team of ACSC/Home Affairs cyber expertise and industry AI security specialists, housed in ACSC or DTA, working with agencies moving AI from proof of concept to production. Initial priority: AI deployments on Systems of Government Significance, high-risk use cases identified through the risk-tiered classification (Rec 8), and GovAI Chat as it rolls out in 2026. This Lab is transitional: it exists to bridge the gap while ISM AI controls (Rec 6),

> IRAP evolution (Rec 7), and the workforce pipeline (Recs 16–17) mature. It directly feeds into how ACSC can evolve its technical Commonwealth cyber programs for the AI era. *(ACSC / DTA)* [S8]

The institutional home is deliberately listed as either ACSC or DTA. ACSC brings operational proximity to threat intelligence and ISM ownership; DTA brings deployment proximity and agency adoption authority. The choice depends on whether the capability is primarily a security function or an adoption-enablement function. For CISOs, the immediate value is access to specialist AI security assessment capability that agencies cannot build individually, and operational findings that feed directly into executive risk reporting on AI security alongside conventional ICT risk.

As existing programs evolve, HBS extends to capture AI-relevant signals at the endpoint level (distinct from the prompt and deployment-layer logging addressed by Rec 14), CHIPs extends to AI vulnerability scanning, and CTIS extends to AI threat indicators (Rec 15). Independent PSPF compliance verification and the Posture Report (Recs 3, 4) provide the emerging assurance layer.


AI Security Lab model
Enduring ACSC programs (evolving with Lab input)
HBS → AI-relevant endpoint signals
CHIPs → AI vulnerability scanning
CTIS → AI-specific threat indicators
ISM → AI control development
AI security skills development (Recs 16–17)
Findings, direction, AI context
Transitional AI Security Lab (ACSC / DTA)
Foundation model security assessment
Common component evaluation
Red-team methodology and tooling
Agent security testing (incl. MCP/A2A)
SoGS-specific risk and control assessments
Direct agency deployment support
Direct deployment support
Incidents, patterns, lessons
Agency level: CISOs
Deployment-specific security
RAG configuration and access controls
Prompt management and output handling
Fine-tuning decisions
Agency SIEM/SOC integration (Rec 14)
AI-specific pen testing (Rec 13)
Assurance: Independent PSPF compliance verification (Rec 3) + AI Impact Assessment assurance (Rec 4)

AI Security Lab model

A three-tier model applies. The Lab itself conducts foundation model security assessment, common component evaluation, red-team methodology development, agent security testing (including emerging protocols like MCP and A2A), and SoGS-specific risk and control assessments. It provides direction and AI-specific context to ACSC's enduring programs (HBS, CHIPs, CTIS, ISM development) so those capabilities evolve to address AI threats without being absorbed into a temporary structure. And it works directly alongside agency CISOs through AI deployment transitions: conducting deployment-specific risk assessments, advising on control implementation, and building the practical AI security knowledge that agencies currently lack. This direct engagement serves a dual purpose: it provides agencies with expert support they cannot yet recruit, and it generates the operational experience base that feeds into AI security workforce development (Recs 16–17) and ACSC's own program evolution. The mechanism — whether through structured secondments, training rotations, or operational case documentation — requires design as part of Lab establishment. For genuinely new risks where no ISM control exists (model provenance, prompt injection, AI-specific logging, RAG pipeline security), the Lab develops assessment criteria; for AI-context gaps where existing controls need AI context (identity management, supply chain), it translates requirements into practical agency guidance through direct work with agency CISOs implementing those controls. Agency-level responsibility covers deployment-specific security: fine-tuning decisions, RAG configuration, access controls, prompt management, and output handling.

8.3 What agencies and CISOs can do now

ASD is building AI controls into the ISM and DTA has published proof-of-concept-to-production guidance [132]. Mature international frameworks exist now. The ISM's risk-based architecture supports agencies applying controls beyond the minimum prescribed. CISOs should not wait.

AI asset inventory. Extend beyond standalone AI to include AI embedded in vendor products: Copilot features in M365 tenants, AI in ServiceNow, Splunk, HR and finance platforms. Answer: what AI is running, on what data, with what access, under whose accountability?

AI-specific third-party risk assessments. Where does the vendor's model process your data, what training data was used, what controls exist against prompt injection, does the vendor retain or learn from your inputs?

Apply available frameworks now. NIST AI RMF, OWASP Top 10 for LLMs, MITRE ATLAS, and ASD's own eight AI security publications are mature, free, and referenced in Australian government publications [76][79][80][11]. Map agency AI deployments against them today.

AI-specific penetration testing. AI red-teaming extends existing penetration testing to include AI-specific test cases. What makes it different from conventional is that the attack surface is different. Prompt injection is not a network vulnerability; it exploits the natural language interface itself. System-level testing (how the AI interacts with agency data, APIs, and access controls) matters more than testing the model in isolation. Microsoft's AI Red Team has tested more than 100 generative AI products since 2018, finding that simple attacks are often most effective and that system-level testing matters more than testing models in isolation [96]. Open-source tooling exists: PyRIT, ATLAS, Inspect [97][80]. Unit 42 has documented 22 distinct prompt injection techniques actively exploited in production, including attacks targeting data destruction and unauthorised financial transactions [155], reinforcing why this matters.

AI-specific logging. The ISM March 2026 update strengthened event logging. DET-01 now specifies protection of logs against modification and deletion, and DET-04 introduces baselined high-risk access activities for anomaly detection [131]. These are welcome foundations. But the ISM logging controls contain no explicit coverage of AI-specific events. ISM-0585 records user, process, filename, and event description. None maps to prompt interactions, model lifecycle events, RAG query patterns, or agent action chains. An agency fully ISM-compliant under March 2026 has zero visibility into AI-specific events.

AI threat intelligence sharing. AI-related threats can be shared through CTIS now using existing STIX 2.1 object types: attack patterns describing prompt injection techniques, indicators capturing malicious model hashes, and malware objects for AI-specific exploits.

The intelligence is shareable even though STIX 2.1 does not yet have AI-specific object types as a standard. AI-specific extensions are emerging through community efforts, and ACSC's role in driving standardisation through the CTIS platform is where Recommendation 15 sits. In the meantime, CISOs should ensure their teams are reporting AI-related incidents and threat observations through existing CTIS channels.

AI-specific incident response playbooks. Existing incident response plans do not cover AI-specific scenarios. CISOs should develop or adapt playbooks addressing compromised models (detection, isolation, rollback), chatbot data spills (classification of exposed data, notification triggers, prompt chain forensics), and adversarial prompt injection (distinguishing malicious from benign usage in logs). As part of that development, review access controls for AI systems: privileged access inheritance into AI environments, AI service account permissions, and whether existing break-glass procedures account for AI systems operating autonomously.

> **Recommendation 13 (Now):** Extend penetration testing to AI attack surfaces. For AI processing sensitive data, make this a condition of operational approval. *(CISOs)* [S8]
>
> **Recommendation 14 (Now):** Mandate AI-specific logging and audit trails for all government AI systems, integrated into existing SIEM/SOC operations. *(ACSC & CISOs)* [S8]
>
> **Recommendation 15 (Next):** Extend CTIS to AI-specific threat types as a centralised ACSC responsibility. This positions ACSC as the national clearing house for AI threat intelligence, enabling machine-speed sharing of AI-specific indicators across agencies. *(ACSC)* [S8]

9. Current Cyber Security Workforce

Every recommendation in this paper requires people to implement it. The workforce data is unambiguous: the skilled AI security people needed do not exist in sufficient numbers.

9.1 The numbers

More than half of APS agencies report critical cyber security skills shortages [69]. The DTA estimates a digital talent shortfall exceeding 8,000 people by 2030 [70]. Over 20% of the ICT and Digital Solutions workforce is approaching retirement age [69]. Agencies cite affordability as one factor preventing recruitment, but the structural barriers extend well beyond salary. Sixty per cent of APS technology jobs are concentrated in the ACT, drawing from just 4.6% of the national ICT workforce; the DTA projects it would take 132 years to rebalance this geographic concentration at the current pace [70]. Average recruitment time to merit pool selection is 69 days, compared to 33 days in the private sector [72].

These are not new findings. They describe a workforce system that was already struggling to meet the demands of conventional cyber security before AI entered the picture. AI adoption adds new skill requirements (model security, AI red-teaming, AI-specific incident response, prompt injection defence, AI supply chain assurance) to a workforce base that is already insufficient for the pre-AI mandate.

9.2 The pipeline gap

The Digital Traineeship program, targeting 1,000 participants over four years [69], is building a generalist digital pipeline. But pipeline programs are producing generalist cyber professionals, not the AI security specialists the next phase demands. The gap between generalist cyber capability and AI security specialism is the gap between understanding network segmentation and understanding how to assess a RAG pipeline for data exfiltration risk, or how to red-team a large language model deployment.

The ASD Cyber Skills Framework, the reference document defining cyber security roles and career pathways for the APS, was last updated in September 2020 [12]. It contains no AI security competencies, no AI-specific roles, no mention of adversarial ML, AI supply chain risk, or model assurance. The Australian Cyber Workforce Playbook (October 2025) acknowledges AI will reshape cyber security roles [130] but does not define AI security specialist competencies. Like what was experienced with cloud security over the past ten years, having defined roles initially helps build depth and specialisation. Eventually the rest of the cyber workforce catches up, but the initial pool of experts needs to be created first.

9.3 The world has moved

In December 2025, NIST released an updated NICE Framework (v2.1.0) that includes a dedicated AI Security Competency Area (NF-COM-002) [128]. This is the framework that underpins US federal cyber workforce planning. Within a twelve-month window, every major certification body (ISC2, SANS/GIAC, CompTIA, ISACA) launched AI security credentials [133][134][135][136]. ISC2's 2025 Workforce Study ranked AI as the number one skill needed, cited by 41% of respondents [133]. The UK has identified 14 specialist AI security firms and approximately 9,740 AI security-linked roles [129]. Even with dedicated specialist firms, the UK is worried about capacity. Australia, without equivalent specialist firms or role definitions, has further to travel.

International AI security credentials do not cover the Australian policy and controls landscape: ISM, PSPF, Essential Eight, and IRAP. That gap is where targeted domestic investment is needed.

The shift toward insourcing has built sovereign capacity but reduced the diversity of experience within government cyber teams. AI security expertise currently sits predominantly in the private sector outside of Canberra, where earlier adoption has produced practical knowledge that government teams have not yet had the opportunity to develop. Agencies should draw on that expertise through their cyber uplift and AI programs. Not as consulting dependency, but as recognition that AI security lessons from industry need to transfer into government.

> **Recommendation 16 (Next):** Create AI security specialist roles within the ASD Cyber Skills Framework, incorporating data governance literacy as a core competency. Not to do governance work, but to identify where governance gaps create security vulnerabilities. Integrate into DTA's APS Cyber Workforce development structures. Defined roles create the recruitment targets, career pathways, and training investment cases that the current generic cyber workforce framework cannot support. *(DTA / ACSC)* [S9]

> **Recommendation 17 (Next):** Accelerate AI security skills development through existing training mechanisms with Australian contextualisation (ISM, PSPF, IRAP mapping) that international credentials alone cannot provide. *(DTA / ACSC)* [S9]

9.4 The binding constraint

The workforce gap is the binding constraint on everything else in this paper. The ISM can be extended with AI security controls, but someone needs to implement and assess compliance against them. IRAP can evolve to cover AI workloads, but assessors need AI security competencies to conduct those assessments. AI-specific logging can be mandated, but SOC analysts need to know what AI-specific anomalies look like. Penetration testing can be extended to AI attack surfaces, but testers need adversarial machine learning skills. Governance frameworks require people who understand both the governance architecture and the technical reality of AI systems.

The cloud security parallel is instructive. Organisations that invested early in dedicated cloud security roles built institutional knowledge that accelerated their broader workforce. Organisations that waited paid a compounding cost in incidents, retrofit, and lost time. AI security is following the same pattern, faster. Creating specialist roles and building sovereign training pathways will not close the gap overnight. But every AI deployment that goes live without the people needed to secure it widens the gap further.

10. Conclusion

The evidence assembled in this paper leads to a single conclusion. Australia’s public sector cyber security foundations are not ready for AI, and AI is arriving regardless. The policy architecture has improved markedly. The frameworks exist. The international reference models are mature. What is missing is the implementation capacity: the modernisation investment to move off legacy IT that constrains security posture, the dedicated funding and workforce to secure what agencies operate, and the structural efficiency to stop duplicating capability across 190+ entities. The gap between what has been mandated and what has been resourced is already the defining constraint.

AI does not create the security problems documented here. It amplifies them, accelerates them, and makes them harder to detect. Every unresolved privileged access failure, every incomplete asset inventory, every self-assessment that overstates reality becomes a larger risk when AI systems operate on top of it. Low data governance maturity compounds the problem: agencies that cannot govern their data cannot secure AI systems that transform it. The deployment layer where AI meets agency data, integrations, and access controls remains unassessed by any assurance mechanism. And the self-assessment gap documented across a decade of cyber compliance reporting is on course to repeat in AI governance, where AI Impact Assessments and Responsible AI Maturity Assessments, based on publicly available information, rely on the same self-assessment model. Independent assurance, the correction Home Affairs is now piloting through the PSPF Assurance Capability [42], must transfer to AI governance before the gap opens.

The seventeen recommendations in this paper are designed to be proportionate and achievable within existing institutional structures. None requires new legislation. None requires new agencies. They require the decision to resource what has been mandated: modernisation that delivers security improvement as a byproduct, dedicated investment in the people and tools to operate modern security, and efficient structural models for shared capability. Cyber security must be treated as a precondition for the AI-driven productivity gains the public sector is counting on.

This is not an argument to halt AI adoption until every cyber security gap is closed. That would mean waiting indefinitely. It is an argument that agencies need a higher security threshold for high-impact AI use cases, independent assurance for production deployments processing sensitive data, and tighter controls where AI interacts with sensitive government data. ASD is doing the right thing: engaging with industry and the assessor community, building AI controls into the ISM progressively, consulting on what to build next. That process produces better controls. It also operates on a cadence (quarterly updates, stakeholder consultation, agency budget cycles) that is structurally slower than AI adoption. The transitional assessment capability proposed in Recommendation 2 exists specifically to bridge that pace gap while the formal build continues. And the accountability mechanism itself, which has evolved to focus on a subset of eight controls when the threat landscape was narrower, must now broaden to capture the reforms already delivered. Getting the security foundations right enables AI adoption; deferring them guarantees the amplification effect will compound.

The window in which these foundations can be secured before AI scales beyond experimentation is narrow, and it is closing.

Appendix A: AI Security Threats

The following provides examples of the types of security threats that the AI attack surface introduces. These are documented incidents and demonstrated vulnerabilities. Each is explained in the context of why it is relevant to Australian government agencies deploying AI into environments where the security foundations documented in this paper remain incomplete.

Prompt injection: the front door is open

Any AI system that accepts natural language input is a potential attack surface. Prompt injection works by embedding malicious instructions within inputs that an AI system processes as legitimate, overriding intended behaviour to extract data, bypass safety controls, or manipulate outputs. In March 2026, Unit 42 confirmed active web-based indirect prompt injection on live websites, documenting 22 distinct attacker payload techniques with intents including data destruction, AI content moderation bypass, and unauthorised financial transactions [155]. The first documented case of prompt injection bypassing an AI-based ad review system involved a single scam page containing 24 separate injection attempts using layered delivery techniques [155]. Cisco and the University of Pennsylvania tested DeepSeek R1 using automated adversarial techniques across 50 randomly sampled HarmBench prompts and achieved a 100% jailbreak success rate; other methodologies found 58–96% for the same model [111]. Defences are improving (input filtering, output monitoring, system prompt hardening), but no current mitigation is reliable against an adaptive attacker with sustained access. Prompt injection is the number one vulnerability in OWASP’s Top 10 for LLM Applications 2025 [79].

Why this matters for government: GovAI Chat and any agency AI tool that accepts natural language input from public servants or citizens is exposed to this attack vector. An AI system processing citizen queries, summarising casework, or drafting correspondence can be manipulated to disclose information from its context window, bypass output restrictions, or produce misleading content. The deployment-layer controls discussed in Section 7.5 (prompt-layer protection, output filtering, logging of prompt interactions) are the defences that the ISM does not yet mandate.

Data poisoning: corruption you cannot see

Data poisoning attacks compromise AI systems at the training stage, embedding behaviours that activate under specific conditions while remaining undetectable during normal operation. Anthropic’s “Sleeper Agents” research demonstrated that backdoors inserted during training can persist through standard safety training processes designed to remove them [105]. A Nature Medicine study made the practical implications concrete: poisoning just 0.001% of training tokens in a medical language model produced a 4.8% increase in harmful outputs, nearly undetectable by human evaluators [106].

Why this matters for government: Agencies relying on AI-assisted decision-making in healthcare, fraud detection, visa processing, or citizen services face a specific risk: a poisoned model does not announce itself. It produces outputs that look correct but are systematically compromised. Where agencies fine-tune models on their own data or use RAG pipelines drawing from agency document holdings, the integrity of that data

becomes a direct security input. The data governance maturity score of 2.02 out of 5 documented in Section 5.2 means most agencies lack the governance foundations to assure training data integrity.

Model supply chain compromise

JFrog Security discovered over 100 malicious AI models on Hugging Face containing hidden backdoors [107]. Palo Alto Networks demonstrated namespace hijacking attacks on model repositories referenced by Google Vertex AI and Azure AI Foundry [108]. Model provenance is harder to verify than software provenance, and the consequences of a compromised model in a government decision system are qualitatively different from a compromised software library.

Why this matters for government: Agencies deploying open-source models (Pattern 2 in Section 4.2) or using vendor platforms that source models from public repositories inherit these supply chain risks. The absence of mandatory SBOM requirements documented in Section 7.7 means agencies have no standardised mechanism to trace model provenance, verify training data sources, or identify downstream dependencies. A compromised model embedded in a vendor product (Pattern 3 or 4) may not be visible to the agency at all.

Nation-state targeting of AI systems

The EchoLeak vulnerability (CVE-2025-32711, CVSS 9.3) demonstrated zero-click data exfiltration from Microsoft 365 Copilot. A malicious document or email processed by Copilot could trigger exfiltration through classifier evasion and auto-fetched image exploitation [109]. Google's Threat Intelligence Group reported that government-backed actors from North Korea, Russia, China, and Iran used Gemini for reconnaissance, targeting research, and phishing lure development — treating commercial AI as a productivity tool for intelligence operations [110]. Separately, GTIG detected and disrupted model extraction campaigns exceeding 100,000 prompts directed at Gemini from private-sector entities and researchers globally, attempting to replicate the model's reasoning capabilities [110].

Why this matters for government: The relevance is direct. Agencies deploying Copilot within M365 tenants are deploying the same platform the EchoLeak vulnerability targeted, and approximately 75% did not receive independent assurance covering all their cloud computing services [21]. AI systems processing sensitive government data are high-value intelligence targets. The combination of nation-state capability and the detection speed documented in Section 5.3 (74% of government breaches take more than 30 days to identify) means a sophisticated adversary exploiting an AI system has a substantial operational window before detection.

Insider risk and shadow AI

Samsung's experience is instructive: within 20 days of engineers gaining access to ChatGPT, three separate incidents occurred where employees input proprietary source code [112]. Industry data shows 68% of enterprise employees access public AI tools through personal accounts, with 57% admitting to entering sensitive information [141].

Why this matters for government: Shadow AI intersects with foreign interference and FOCI risks that the PSPF 2025 strengthened reporting requirements to address [24]. An

authorised user with access to an AI system processing sensitive data who is subject to foreign influence can extract and exfiltrate information at a speed pre-AI controls were not designed to detect. The cultural pattern is predictable: when AI tools make public servants more productive, they will use them whether approved or not. Operational pressure consistently overrides security compliance when controls are seen as barriers to getting work done. Shadow AI follows exactly the pattern of shadow IT, except it processes data faster, at greater scale, and with less visibility.

AI infrastructure as a target

AI systems run on specialised infrastructure: ML frameworks (Ray, PyTorch, TensorFlow), model serving platforms, vector databases, GPU clusters, and orchestration layers. This infrastructure is itself an attack surface. The ShadowRay campaign (CVE-2023-48022, CVSS 9.8) exploited a critical vulnerability in the Ray AI framework, compromising hundreds of companies' AI infrastructure. Attackers gained access to AI models, training data, API tokens, and cloud credentials simultaneously. A single compromised server contained 240GB of source code, AI models, and API tokens [116].

Why this matters for government: As agencies move from experimentation to production, they deploy AI infrastructure that sits alongside conventional IT systems but is often managed by data science teams rather than security operations. The asset inventory blindness documented in Section 5.2 extends to AI infrastructure: agencies that cannot account for their conventional server estate are unlikely to know what ML frameworks, model serving platforms, or vector databases are running in their environments. This is compounded by Pattern 4 deployments (Section 4.2) where AI features activate through vendor product updates without explicit agency decision. A compromised AI infrastructure component gives an attacker access not just to one system but to the models, data, and credentials that flow through it. The concentration risk is qualitatively different from conventional IT: compromising a single Ray cluster or model serving endpoint can expose the agency's entire AI capability simultaneously.

What this means

MITRE ATLAS catalogues 15 tactics, 66 techniques, and 33 real-world case studies specific to AI (as of October 2025) [80]. Approximately 70% of ATLAS mitigations map to existing security controls [115]. The defence relies on getting the fundamentals documented in this paper right, then building AI-specific capabilities on top. Agencies with mature security implementations have a foundation to build on. The audit evidence in Section 5 suggests that foundation is not solid, and the threats documented above will land on whatever security posture agencies have when AI goes live.

Appendix B: Acronyms and Definitions

Acronym	Definition
ACSC	Australian Cyber Security Centre (part of ASD)
AI	Artificial Intelligence
AIA	AI Impact Assessment
AISI	Australian AI Safety Institute
ANAO	Australian National Audit Office
API	Application Programming Interface
APS	Australian Public Service
APSC	Australian Public Service Commission
ASD	Australian Signals Directorate
ATO	Authority to Operate
ATLAS	Adversarial Threat Landscape for AI Systems (MITRE)
AUPDNS	Australian Protective Domain Name Service (ASD program)
AWS	Amazon Web Services
BOM	Bill of Materials
cATO	Continuous Authority to Operate
BOM	Bill of Materials (AI context: model provenance and dependency documentation)
CHIPs	Cyber Hygiene Improvement Programs (ASD program)
CI/CD	Continuous Integration / Continuous Deployment
CISA	Cybersecurity and Infrastructure Security Agency (US)
CISO	Chief Information Security Officer
COSAiS	Cybersecurity of AI Systems (NIST project)
CSP	Cloud Service Provider
CTIS	Cyber Threat Intelligence Sharing (ASD program)
CVE	Common Vulnerabilities and Exposures
CVSS	Common Vulnerability Scoring System
DFAT	Department of Foreign Affairs and Trade
DISR	Department of Industry, Science and Resources
DNS	Domain Name System
DTA	Digital Transformation Agency
E8	Essential Eight
EO	Executive Order (US)
ETSI	European Telecommunications Standards Institute
FedRAMP	Federal Risk and Authorization Management Program (US)
FOCI	Foreign Ownership, Control or Influence
GenAI	Generative Artificial Intelligence

Acronym	Definition
GSPN	Government Security Posture Network (ASD program)
GTG	Google Threat Group
HBS	Host Based Sensor (ASD program)
HCF	Hosting Certification Framework
IDAM	Identity and Access Management
IRAP	Information Security Registered Assessors Program
ISM	Information Security Manual (ASD)
ISO/IEC	International Organization for Standardization / International Electrotechnical Commission
LLM	Large Language Model
MDA	Modern Defensible Architecture (ASD)
MFA	Multi-Factor Authentication
ML	Machine Learning; also Maturity Level (context dependent)
ML2	Maturity Level 2 (Essential Eight)
NCSC	National Cyber Security Centre (UK)
NDB	Notifiable Data Breaches (scheme)
NDAA	National Defense Authorization Act (US)
NICE	National Initiative for Cybersecurity Education (US)
NIST	National Institute of Standards and Technology (US)
OAIC	Office of the Australian Information Commissioner
OMB	Office of Management and Budget (US)
OT	Operational Technology
OWASP	Open Worldwide Application Security Project
PQC	Post-Quantum Cryptography
PSPF	Protective Security Policy Framework
PyRIT	Python Risk Identification Tool (Microsoft)
RAG	Retrieval-Augmented Generation
REDSPICE	Resilience, Effects, Defence, Space, Intelligence, Cyber, Enablers (ASD program)
SAFE-AI	Securing AI Frameworks and Environments (MITRE)
SAIF	Secure AI Framework (Google)
SBOM	Software Bill of Materials
SIEM	Security Information and Event Management
SOCI Act	Security of Critical Infrastructure Act 2018
SOC	Security Operations Centre; also Service Organization Controls (context dependent)
SoGS	Systems of Government Significance

Acronym	Definition
SoNS	Systems of National Significance
STIX	Structured Threat Information Expression
TAXII	Trusted Automated Exchange of Intelligence Information

Key terms

Agentic AI: AI systems that chain tool calls, persist across sessions, take autonomous actions, and interact with other systems — as distinct from chat-based AI where a human initiates each interaction.

AI deployment layer: The combination of model, agency data, integrations, access controls, prompt chains, RAG pipelines, and output handling that constitutes an agency's actual AI deployment — distinct from the underlying cloud infrastructure and the AI model itself.

Authority to Operate (ATO): A formal decision by an authorising officer that a system's security risks are acceptable and the system may operate. In Australia, ATO decisions are agency-led and typically point-in-time.

Continuous Authority to Operate (cATO): An approach to authority to operate where security assurance is maintained through continuous monitoring, automated testing, and ongoing risk assessment rather than periodic point-in-time assessments.

Essential Eight Maturity Level 2 (E8 ML2): The mandatory baseline level of implementation for the ASD Essential Eight mitigation strategies, required for all non-corporate Commonwealth entities since July 2022.

ISM AI security controls: AI-specific controls being progressively embedded into existing ISM control families, defining outcome-oriented security requirements for AI systems.

Prompt injection: An attack technique where malicious instructions are embedded within inputs that an AI system processes as legitimate, overriding intended behaviour to extract data, bypass safety controls, or manipulate outputs.

RAG (Retrieval-Augmented Generation): A technique where an AI model retrieves information from external data sources (such as agency documents) to augment its responses, creating a direct connection between the model and agency data holdings.

Shadow AI: The use of unapproved AI tools and services by employees outside official IT governance and security controls, analogous to shadow IT.

References

Public URLs verified where available as of March 2026. Sources marked with an asterisk () require subscription or registration. Some institutional sources cited by title where public URLs are not available.*

Australian Government — Australian Signals Directorate (ASD)

[1] ASD, *Commonwealth Cyber Security Posture in 2025*, cyber.gov.au, February 2026. https://www.cyber.gov.au/about-us/view-all-content/reports-and-statistics/the-commonwealth-cyber-security-posture-in-2025

[2] ASD, *Commonwealth Cyber Security Posture in 2024*, cyber.gov.au, 2024. https://www.cyber.gov.au/about-us/view-all-content/reports-and-statistics/commonwealth-cyber-security-posture-2024

[3] ASD, *Commonwealth Cyber Security Posture in 2023*, cyber.gov.au, 2023. https://www.cyber.gov.au/about-us/view-all-content/reports-and-statistics/commonwealth-cyber-security-posture-2023

[4] ASD, *Annual Cyber Threat Report 2024–25*, cyber.gov.au, October 2025. https://www.cyber.gov.au/about-us/view-all-content/reports-and-statistics/annual-cyber-threat-report-2024-2025

[5] ASD, *Annual Cyber Threat Report 2023–24*, cyber.gov.au, 2024. https://www.cyber.gov.au/about-us/view-all-content/reports-and-statistics/annual-cyber-threat-report-2023-2024

[6] ASD, *Information Security Manual*, cyber.gov.au, March 2026 (quarterly updates). https://www.cyber.gov.au/business-government/asds-cyber-security-frameworks/ism

[7] ASD, *Modern Defensible Architecture Foundations*, cyber.gov.au, October 2025. https://www.cyber.gov.au/business-government/secure-design/secure-by-design/modern-defensible-architecture/foundations-for-modern-defensible-architecture

[8] ASD, *Essential Eight Maturity Model*, cyber.gov.au, November 2023. https://www.cyber.gov.au/business-government/asds-cyber-security-frameworks/essential-eight/essential-eight-maturity-model

[9] ASD, *Strategies to Mitigate Cyber Security Incidents*, cyber.gov.au, 2017. https://www.cyber.gov.au/resources-business-and-government/essential-cybersecurity/strategies-mitigate-cybersecurity-incidents/strategies-mitigate-cybersecurity-incidents

[11] ASD, AI Security Guidance Suite (eight publications covering development, deployment, data security, supply chain, OT integration, general engagement, and small business guidance), cyber.gov.au, 2024–2026. https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence

[12] ASD, *Cyber Skills Framework Version 2*, cyber.gov.au, September 2020. https://www.cyber.gov.au/business-government/protecting-business-leaders/cyber-security-for-business-leaders/cyber-skills-framework

[13] ASD, CISA, NCSC et al., *Guidelines for Secure AI System Development*, November 2023. https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence/guidelines-for-secure-ai-system-development

[14] ASD, CISA et al., *Deploying AI Systems Securely*, April 2024. https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence/deploying-ai-systems-securely

[15] ASD, *Annual Report 2024–25*, October 2025. https://www.asd.gov.au/about/accountability-governance/publications/asd-annual-report-2024-25

[131] ASD, *Information Security Manual March 2026 Changes*, cyber.gov.au, 17 March 2026. https://www.cyber.gov.au/business-government/asds-cyber-security-frameworks/ism

[144] ASD and AICD, *Cyber Security Priorities for Boards of Directors 2025–26*, cyber.gov.au, October 2025. https://www.cyber.gov.au/business-government/protecting-business-leaders/cyber-security-for-business-leaders/cyber-security-priorities-for-boards-of-directors-2025-26

Australian Government — Australian National Audit Office (ANAO)

[16] ANAO, *AG Report No. 38 of 2023–24: Management of Cyber Security Incidents*, June 2024. https://www.anao.gov.au/work/performance-audit/management-cyber-security-incidents

[18] ANAO, *AG Report No. 26 of 2022–23: Interim Report on Key Financial Controls*, 2023. https://www.anao.gov.au/work/financial-statement-audit/interim-report-key-financial-controls-major-entities-2022-23

[19] ANAO, *AG Report No. 1 of 2019–20: Cyber Resilience*, 2019. https://www.anao.gov.au/work/performance-audit/cyber-resilience-government-business-enterprises-and-corporate-commonwealth-entities

[20] ANAO, *AG Report No. 9 of 2022–23: Management of Cyber Security Supply Chain Risks*, 2022. https://www.anao.gov.au/work/performance-audit/management-cyber-security-supply-chain-risks

[21] ANAO, *AG Report No. 22 of 2024–25: Audits of the Financial Statements of Australian Government Entities for the Period Ended 30 June 2024*, February 2025. https://www.anao.gov.au/work/financial-statement-audit/audits-of-the-financial-statements-of-australian-government-entities-the-period-ended-30-june-2024

[138] ANAO, *Insights: Audit Lessons — Governance of Data*, anao.gov.au, June 2025. https://www.anao.gov.au/work/insights/governance-of-data

Australian Government — Policy Instruments

[24] Department of Home Affairs, *Protective Security Policy Framework Release 2025*, protectivesecurity.gov.au, 24 July 2025. https://www.protectivesecurity.gov.au/publications-library/pspf-annual-release-2025

[27] PSPF Direction 003-2024 (ASD Cyber Security Partnership), 2024. https://www.protectivesecurity.gov.au/publications-library/direction-003-2024-supporting-visibility-cyber-threat

[28] Department of Home Affairs, *Gateway Security Standard 2025*, 24 July 2025. https://www.protectivesecurity.gov.au/publications-library/australian-government-gateway-security-standard-2025

[29] Department of Home Affairs, *Systems of Government Significance Standard*, 24 July 2025. https://www.protectivesecurity.gov.au/pspf-annual-release/pspf-standards

[147] Minister for Home Affairs (Tony Burke), Cyber Security Uplift for Systems of Government Significance, minister.homeaffairs.gov.au, July 2025. https://minister.homeaffairs.gov.au/TonyBurke/Pages/cyber-security-uplift-for-systems-of-government-significance.aspx

[30] Department of Home Affairs, *2023–2030 Australian Cyber Security Strategy*, November 2023. https://www.homeaffairs.gov.au/about-us/our-portfolios/cyber-security/strategy/2023-2030-australian-cyber-security-strategy

[31] DTA, *Policy for the Responsible Use of AI in Government v2.0*, digital.gov.au, 15 December 2025. https://www.digital.gov.au/ai/ai-in-government-policy

[33] Department of Finance, *APS AI Plan 2025*, digital.gov.au, 12 November 2025. https://www.digital.gov.au/policy/ai/australian-public-service-ai-plan-2025/foreword

[34] Department of Finance, media release re GovAI Chat, 19 December 2025. https://www.finance.gov.au/about-us/news/2025/introducing-aps-ai-plan

[35] DISR, *National AI Plan*, 2 December 2025. https://www.industry.gov.au/publications/national-ai-plan

[42] Department of Home Affairs, *PSPF Assessment Report 2024–25*, protectivesecurity.gov.au. https://www.protectivesecurity.gov.au/publications-library/protective-security-policy-framework-pspf-assessment-report-2024-25

[132] DTA, *Guidance for AI proof of concept to scale*, digital.gov.au, March 2026. https://www.digital.gov.au/policy/ai/AI-POC-to-scale

[137] Department of Finance, *Australian Public Service Data Maturity Report 2024*, finance.gov.au, March 2025. https://www.finance.gov.au/sites/default/files/2025-03/2024_Data-Maturity-Report.pdf

Australian Government — Legislation, Budget and Other

[45] 2024–25 Federal Budget Paper No. 2. https://budget.gov.au/content/bp2/download/bp2_2024-25.pdf

[46] 2022–23 Federal Budget Papers (REDSPICE $9.9B). https://archive.budget.gov.au/2022-23/

[47] 2025–26 Defence Portfolio Budget Statements, Table 4a. https://www.defence.gov.au/about/accessing-information/budgets/budget-2025-26

Office of the Australian Information Commissioner (OAIC)

[48] OAIC, *Notifiable Data Breaches Report July–December 2024*. https://www.oaic.gov.au/privacy/notifiable-data-breaches/notifiable-data-breaches-publications/notifiable-data-breaches-report-july-to-december-2024

[50] OAIC, *Notifiable Data Breaches Report July–December 2023*. https://www.oaic.gov.au/privacy/notifiable-data-breaches/notifiable-data-breaches-publications/notifiable-data-breaches-report-july-to-december-2023

[51] OAIC, blog post, 13 May 2025 (reporting on full year 2024 NDB data). https://www.oaic.gov.au/news/blog/latest-notifiable-data-breaches-statistics-for-july-to-december-2024

[52] OAIC, blog post, 13 May 2025 (reporting on H2 2024 NDB data). https://www.oaic.gov.au/news/blog/latest-notifiable-data-breaches-statistics-for-july-to-december-2024

APS Workforce

[69] APSC, *APS Data, Digital and Cyber Workforce Plan 2025–30*, March 2025. https://www.dataanddigital.gov.au/workforce

[70] DTA, *2025 APS Digital Workforce Insights Report*, 2025. https://www.dataanddigital.gov.au/actions-underway/digital-insights-report

[72] APSC, *State of the Service 2021–22* (recruitment times data). https://www.apsc.gov.au/initiatives-and-programs/workforce-information/research-analysis-and-publications/state-service/state-service-report-2021-22

International Sources

[76] NIST, *AI Risk Management Framework (AI RMF 1.0)*, AI 100-1, January 2023. https://doi.org/10.6028/NIST.AI.100-1

[79] OWASP, *Top 10 for LLM Applications 2025*, November 2024. https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/

[80] MITRE, *ATLAS (Adversarial Threat Landscape for AI Systems)*. https://atlas.mitre.org/

[82] UK NCSC, *Code of Practice for the Cyber Security of AI*, January 2025. https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice

[84] Canada, Treasury Board *Directive on Automated Decision-Making*, April 2019. https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592

[87] US, Executive Order 14110, October 2023. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

[88] US, Executive Order 14179, January 2025 (revoking EO 14110). https://www.federalregister.gov/documents/2025/01/31/2025-02172/removing-barriers-to-american-leadership-in-artificial-intelligence

[89] OMB, Memorandum M-25-21, April 2025. https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf

[145] NIST, *Security and Privacy Controls for Information Systems and Organizations*, SP 800-53 Rev 5, September 2020. Over 1,000 controls across 20 families; mandatory for US federal agencies under FISMA, independently assessed by inspectors general. https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final

[146] UK NCSC, *Cyber Assessment Framework*, v4.0, August 2025. 4 objectives, 14 principles, 39 contributing outcomes; operationalised through the GovAssure assurance scheme (April 2023) with third-party validation. https://www.ncsc.gov.uk/collection/cyber-assessment-framework

[139] NIST, *Control Overlays for Securing AI Systems (COSAiS)* project. Concept paper August 2025; annotated outline (Predictive AI use case) January 2026. https://csrc.nist.gov/projects/cosais

Private Sector and Industry

[96] Microsoft, *AI Red Team whitepaper* ("Lessons From Red Teaming 100 Generative AI Products"), January 2025. https://www.microsoft.com/en-us/security/blog/2025/01/13/3-takeaways-from-red-teaming-100-generative-ai-products/

[97] Microsoft, PyRIT (Python Risk Identification Tool), open source. https://github.com/Azure/PyRIT

[99] AWS, *Generative AI Security Scoping Matrix*. https://aws.amazon.com/ai/security/generative-ai-scoping-matrix/

[101] Mandala Partners (commissioned by Microsoft), government cloud spending report, September 2025. https://mandalapartners.com/reports/unlocking-the-productivity-dividend-of-digital-government

Research and Threat Intelligence

[105] Anthropic, *Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training*, arXiv:2401.05566, January 2024. https://arxiv.org/abs/2401.05566

[106] Alber et al., Medical LLM poisoning study, *Nature Medicine*, 2025. https://doi.org/10.1038/s41591-024-03445-1

[107] JFrog Security, malicious AI models on Hugging Face, February 2024. https://jfrog.com/blog/data-scientists-targeted-by-malicious-hugging-face-ml-models-with-silent-backdoor/

[108] Unit 42 (Palo Alto Networks), namespace hijacking affecting Google Vertex AI and Azure AI Foundry, 2025. https://unit42.paloaltonetworks.com/model-namespace-reuse/

[109] CVE-2025-32711 (EchoLeak), M365 Copilot data exfiltration vulnerability. https://arxiv.org/abs/2509.10540

[110] Google Threat Intelligence Group, government-backed actor misuse of Gemini and model extraction/distillation campaigns (>100,000 prompts), 2025–2026. https://cloud.google.com/blog/topics/threat-intelligence/adversarial-misuse-generative-ai; https://cloud.google.com/blog/topics/threat-intelligence/distillation-experimentation-integration-ai-adversarial-use

[111] Cisco and University of Pennsylvania, DeepSeek R1 jailbreak testing, January 2025. https://blogs.cisco.com/security/evaluating-security-risk-in-deepseek-and-other-frontier-reasoning-models

[112] Samsung ChatGPT data input incident, March 2023. https://techcrunch.com/2023/05/02/samsung-bans-use-of-generative-ai-tools-like-chatgpt-after-april-internal-data-leak/

[115] Vectra AI, MITRE ATLAS mitigations analysis, 2025. https://www.vectra.ai/topics/mitre-atlas

[116] MITRE, Campaign C0045 (ShadowRay), CVE-2023-48022. https://attack.mitre.org/campaigns/C0045/

[140] MITRE, *SAFE-AI: A Framework for Securing AI-Enabled Systems Using NIST SP 800-53*, 2025. Identifies 100 SP 800-53 controls as "potentially AI-affected" across four system elements. https://atlas.mitre.org/pdf-files/SAFEAI_Full_Report.pdf

[141] TELUS Digital, *AI at Work Survey: Enterprise Employees Are Entering Sensitive Data Into AI Assistants More Than You Think*, 26 February 2025. Survey of 1,000 US enterprise employees, January 2025. https://www.telusdigital.com/about/newsroom/telus-digital-survey-reveals-enterprise-employees-use-of-shadow-ai

[142] Anthropic, *Disrupting AI-Enabled Espionage*, anthropic.com, 2025. https://www.anthropic.com/news/disrupting-AI-espionage

[143] White House, *America's AI Action Plan*, July 2025. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

[148] MITRE Center for Threat-Informed Defense, *MITRE ATLAS OpenClaw Investigation*, February 2026. https://ctid.mitre.org/blog/2026/02/09/mitre-atlas-openclaw-investigation/

[149] Knostic Research, Exposing the Unseen: Mapping MCP Servers Across the Internet (1,862 internet-exposed MCP servers, 119 manually verified, zero authentication), July 2025. https://www.knostic.ai/blog/mapping-mcp-servers-study

[150] ANAO, AG Report No. 53 of 2017–18: Cyber Resilience, May 2018. Recommended strengthening processes for verifying the accuracy of entities' PSPF self-assessments; found shortcomings in the E8 Maturity Model that could lead to entities inadvertently overstating compliance. https://www.anao.gov.au/work/performance-audit/cyber-resilience-2017-18

[151] ANAO, AG Report No. 32 of 2020–21: Cyber Security Strategies of Non-Corporate Commonwealth Entities, March 2021. Found two of three entities that reported full implementation of Top Four mitigation strategies had self-assessed inaccurately. https://www.anao.gov.au/work/performance-audit/cyber-security-strategies-non-corporate-commonwealth-entities

[152] DTA, Australian Government AI Technical Standard, Version 1, July 2025. https://www.digital.gov.au/policy/ai/ai-technical-standard

[153] DTA, Pilot AI Assurance Framework and Guidance, digital.gov.au. https://www.digital.gov.au/policy/ai/pilot-ai-assurance-framework/guidance

[154] DTA, Guidance for the AI Impact Assessment Tool, December 2025. https://digital.gov.au/ai/ai-in-government-policy

[155] Unit 42 (Palo Alto Networks), Fooling AI Agents: Web-Based Indirect Prompt Injection Observed in the Wild, 3 March 2026. https://unit42.paloaltonetworks.com/ai-agent-prompt-injection/

[156] OAIC, Notifiable Data Breaches Statistics Dashboard (Jul–Dec 2025, Australian Government, Time Taken to Identify Breaches: 74% >30 days), accessed 24 March 2026. https://www.oaic.gov.au/privacy/notifiable-data-breaches/notifiable-data-breaches-statistics-dashboard

[157] GovTech Singapore, How We Built the AI Guardian Team at GovTech Singapore (self-reported usage figure), Medium, 2025. https://medium.com/aiguardian-govtech/how-we-built-the-ai-guardian-team-at-govtech-singapore-3758cf21004d

Industry Analysis and Media

[117] KPMG, Federal Budget Analysis, 25 March 2025. https://kpmg.com/au/en/home/insights/2025/03/federal-budget-australia.html

[118] CyberDaily, budget analysis, 25 March 2025. https://www.cyberdaily.au/security/11897-federal-budget-2025-cyber-security-loses-out-in-pre-election-budget

[119] Canberra Times, ASD Annual Report/REDSPICE reporting, 7 October 2025. * https://www.canberratimes.com.au/story/9081606/asd-tackles-fraud-reporting-surge-post-redspice-funding-boost/

APS Workforce Development

[128] NIST, *NICE Framework Components v2.1.0*, December 2025. https://www.nist.gov/news-events/news/2025/12/nice-releases-nice-framework-components-v210

[129] DSIT, *Cyber Security Skills in the UK Labour Market 2025*, September 2025; DSIT, *Cyber Security Sectoral Analysis 2025*, March 2025. https://www.gov.uk/government/publications/cyber-security-skills-in-the-uk-labour-market-2025

[130] Department of Home Affairs, *Australian Cyber Workforce Playbook*, October 2025. https://www.homeaffairs.gov.au/cyber-security-subsite/files/australian-cyber-workforce-playbook.pdf

[133] ISC2, *2025 Cybersecurity Workforce Study*, December 2025. AI ranked as the number one skill needed at 41% of respondents. https://www.isc2.org/Insights/2025/12/2025-ISC2-Cybersecurity-Workforce-Study

[134] SANS Institute and GIAC, AI-focused cybersecurity certifications announcement (four certifications covering offensive AI, red team automation, model integrity, and AI-driven operations), 9 September 2025. https://www.giac.org/focus-areas/artificial-intelligence/

[135] CompTIA, SecAI+ certification launch, 17 February 2026. https://www.comptia.org/en-us/certifications/secai/

[136] ISACA, Advanced in AI Security Management (AAISM) certification, August 2025; Advanced in AI Audit (AAIA), May 2025; Advanced in AI Risk (AAIR) beta, December 2025. https://www.isaca.org/credentialing/aaism