

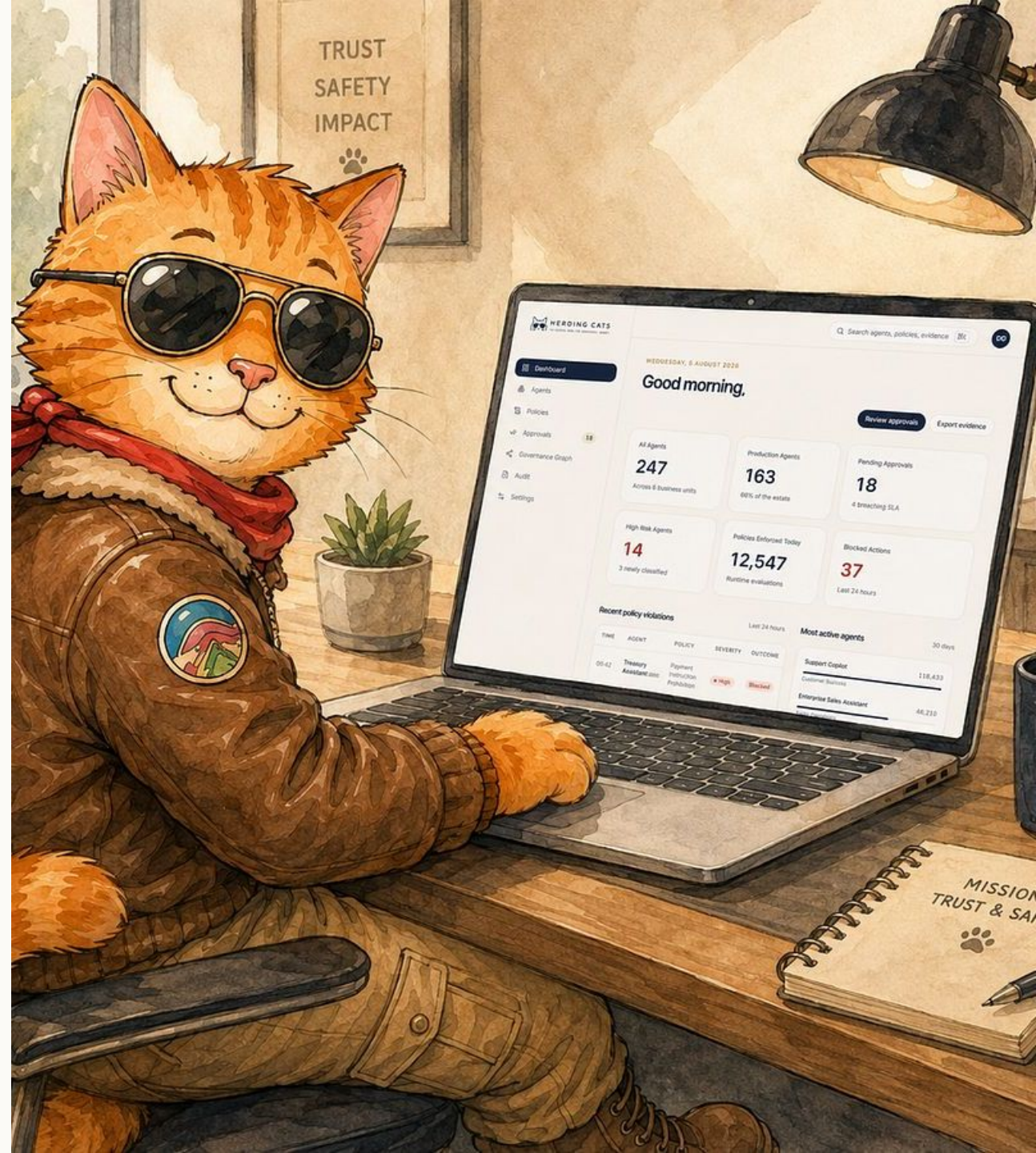


# HERDING CATS

THE CONTROL PLANE FOR ENTERPRISE AI AGENTS

## Australian Information Security Manual (ISM)

The Australian Information Security Manual now includes controls addressing artificial intelligence and agentic AI. This guide highlights the controls most relevant to governing enterprise AI agents, and where Herding Cats can help organisations implement practical technical safeguards.



## Artificial intelligence agent identification

In addition to the general requirement for users to be uniquely identifiable, assigning each AI agent its own unique identity ensures that actions taken by an AI agent can be distinguished from actions taken directly by a user on whose behalf it acts, as well as from actions taken by other AI agents. In doing so, identities assigned to AI agents should be managed in a similar manner to service accounts, including being disabled or removed when no longer required. Where an AI agent is assigned multiple identities, such as across different environments or platforms, all identities should be recorded in an organisation's AI agent register.

**Control: ISM-2133; Revision: 0; Updated: Sep-26; Applicable: NC, OS, P, S, TS; Essential 8: N/A**

Each AI agent is assigned a unique identity that is distinct from the user accounts of personnel and the identities of other AI agents.

## Artificial intelligence agent register

AI agents can be rapidly created and deployed across an organisation's systems, often by personnel operating outside of traditional software deployment processes. Each AI agent may be configured with its own instructions, AI models, tools, permissions, credentials and access to data repositories. Without visibility of the AI agents operating across its systems, an organisation cannot effectively restrict AI agents to the minimum permissions they require, review their continued need, identify unauthorised or abandoned AI agents, or investigate cyber security incidents involving them. As such, an organisation should develop, implement, maintain and regularly verify an AI agent register. An AI agent register may also record additional details, such as the AI models used by each AI agent and their versions, to support change management and cyber supply chain risk management activities.

**Control: ISM-2134; Revision: 0; Updated: Sep-26; Applicable: NC, OS, P, S, TS; Essential 8: N/A**

An AI agent register is developed, implemented, maintained and regularly verified.

**Control: ISM-2135; Revision: 0; Updated: Sep-26; Applicable: NC, OS, P, S, TS; Essential 8: N/A**

An AI agent register contains the following for each AI agent:

- its unique identifier
- its owner and business purpose
- the identities assigned to it
- any user accounts and credentials it uses
- the tools, permissions and data repositories it can access.

## **Excessive Agency**

Excessive agency vulnerabilities allow for damaging actions to be performed in response to unexpected, ambiguous or manipulated output from AI applications, regardless of the cause of the malfunction.

Furthermore, agentic AI applications that can autonomously invoke tools, browse external content, execute code or call other systems introduce risks beyond those of conversational AI applications. In particular, untrusted content retrieved by an agentic AI application, whether via web browsing, email, document processing or retrieval-augmented generation, may contain instructions that the underlying AI model treats as legitimate, allowing malicious actors to indirectly control the agentic AI application's actions. Validation, sanitisation and delimitation can reduce this security risk but cannot be relied upon to make retrieved content trustworthy or eliminate indirect prompt injection. As such, agentic AI applications should treat retrieved content as untrusted throughout processing and prevent it from modifying or overriding system instructions, security policies, access controls, tool permissions or human approval requirements.

**Control: ISM-2092; Revision: 0; Updated: Dec-25;**  
**Applicable: NC, OS, P, S, TS; Essential 8: N/A**

Access control policies are implemented to enforce fine-grained permissions for AI applications.

**Control: ISM-2093; Revision: 0; Updated: Dec-25;**  
**Applicable: NC, OS, P, S, TS; Essential 8: N/A**

Role-based access controls are implemented for AI applications to restrict access to sensitive data.

**Control: ISM-2156; Revision: 0; Updated: Sep-26;  
Applicable: NC, OS, P, S, TS; Essential 8: N/A**

Agentic AI applications are restricted to the minimum set of tools, functions and permissions required for their intended purpose.

**Control: ISM-2157; Revision: 0; Updated: Sep-26;  
Applicable: NC, OS, P, S, TS; Essential 8: N/A**

Tools invoked by agentic AI applications are subject to both the access controls of the invoking user and agent-specific, task-scoped authorisation, with effective permissions limited to the minimum permitted by both.

**Control: ISM-2159; Revision: 0; Updated: Sep-26; Applicable: NC, OS, P, S, TS; Essential 8: N/A**

All tool invocations, external requests and outputs generated by agentic AI applications are centrally logged with sufficient detail to support cyber security incident investigations.

## Still think Agentic AI is irrelevant?

Think again. The ACSC ISM calls for the use of Agentic AI across scenarios including vulnerability assessment, incident response, and software testing. Agentic isn't a future concept. It's becoming part of the cybersecurity playbook.

**Control: ISM-2119; Revision: 0;**  
**Updated: Jun-26; Applicable:**  
**NC, OS, P, S, TS; Essential 8:**  
**N/A**

Suitable AI models are used to augment vulnerability assessments and penetration tests.

**Control: ISM-2117; Revision: 0;**  
**Updated: Jun-26; Applicable:**  
**NC, OS, P, S, TS; Essential 8:**  
**N/A**

Suitable AI models are used to augment the detection of cyber security events and the identification of cyber security incidents.

**Control: ISM-2122; Revision: 0;**  
**Updated: Jun-26; Applicable:**  
**NC, OS, P, S, TS; Essential 8:**  
**N/A**

Suitable AI models are used to augment software security testing.