Australian Government
Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC Australian Cyber Security Centre

Agentic AI Harnesses

The layer above the model

First published: September 2026

Table of contents

Purpose

This publication explains the concept of the Agentic AI harness (referred to hereafter as the harness), the software layer that enables a large language model (LLM) to interact with data, tools and systems to perform agentic tasks. It examines how harness design influences the security, governance, reliability and operational risks of agentic AI systems.

ASD's Careful adoption of agentic AI services provides broader guidance on agentic AI security considerations, including expanded attack surfaces, system complexity and the evolving security landscape. It also outlines the principal risk categories of privilege, design and configuration, behavioural, structural and accountability risks, and provides best-practice guidance for designing, developing, deploying and operating secure agentic AI systems. The harness publication complements that guidance by focusing on the harness as the component organisations can most directly govern, secure and control.

Audience

This publication is intended for organisations adopting or considering agentic AI. It is written for executive decision-makers, chief information security officers and IT leaders.

Executive summary

A harness is the software layer that enables a large language model (LLM) to operate within real-world environments. If you think of the LLM as the brain, then the harness is the body. The LLM takes in information in the form of context, reasons about it and proposes actions to gain more information or complete the task. The harness provides information, as context and memory, it performs actions to access data or use tools, and it can enforce permission and controls on that data and tools. Together, these form an agentic AI system.

Understanding this distinction is important because while some risks stem from the behaviour of the LLM itself, many of the highest-impact organisational risks emerge when the model is connected to enterprise data, systems and tools through a harness. Security, privacy, governance and operational risks are often a consequence of what an agent can access, what actions it can perform, and how it interacts with organisational systems. Some risks, including prompt injection, cannot be reliably addressed within the model alone and require additional controls across the harness, connected systems and organisational governance processes. The harness and its associated integrations, governance mechanisms and operational processes are likely to represent a significant long-term organisational investment. While LLMs will continue to evolve and can be replaced over time, the harness ecosystem, is likely to become the more enduring organisational capability. Organisations should treat the harness as key component of their AI system and apply security and governance controls commensurate with the risk it introduces.

No harness is inherently secure. Organisations should apply established cyber security practices, including least-privilege access, identity and access management, secure design, monitoring, human oversight and integration with incident response processes. Agentic AI should be introduced through a phased approach that defines approved use cases, classifies data, selects appropriate harnesses, applies the controls described in *ASD's* *Careful adoption of agentic AI services*, and validates security controls before broader deployment.

Successful adoption requires matching the model and harness to the task, granting only the access needed to perform that task, applying controls across multiple layers of the system, and maintaining clear human accountability for decisions, actions and outcomes.

Key takeaways

- organisations control the harness, not the LLM
- the harness is where long-term organisational value, governance and investment accumulate.
- the harness and its surrounding technical ecosystem are the primary focus for security and risk management

What is a harness

An agentic AI system comprises a LLM alongside external tools, external data sources, memory and planning workflows. Together, these components enable the system to perceive its environment and, where applicable, take action to achieve its goals.

This publication uses the term 'harness' to describe the components of an agentic AI system other than the LLM itself. The harness is the software layer that connects the LLM to external tools, data sources, memory and planning workflows. It also enforces the system's action and execution privileges and operates the control loop that repeats until a task is complete. In the terms used in *Careful adoption of agentic AI services*, the LLM is the statistical model used to identify what actions to take. The harness supplies the agent's information input, provides its tool or service access, enforces its action and execution privileges, holds its memory and planning workflows, and records its metrics for evaluation.

Identifying the harness as a distinct part of an agentic AI system allows organisations to assess and govern it separately from the LLM. This is important because many of the decisions that determine an agentic AI system's reliability, operating cost and security risk are made in the harness, not in the LLM. In some cases, the harness is developed by the organisation; in others it is supplied as part of a commercial product or service. Regardless of its origin, organisations should understand the harness's capabilities, permissions, integration points and security controls, and assess the risks associated with its use.

The figure below summarises the key differences between an LLM and a harness.

Dimension	LLM	Harness
Core job	Predict the next words or text; reasons over the text in front of it.	Decide what text goes in front of the model, and what to do with the output.
State	Stateless – each interaction starts without memory or previous interactions.	Stateful - persists context, files, memory and progress across turns and sessions.
Actions	Produces tokens only.	Executes tools, edits files, runs commands, calls APIs, browses.
Memory	Only the current context window.	Long-term memory, session logs, scratchpads, files.
Control	Produces outputs but does not directly control any resulting actions.	Permissions, sandboxes, approval gates, hooks, policies.
Consistency	Same input can yield different output.	Adds validation, structured output, retires and error recovery.
Failure mode	Hallucination, wrong answer.	Wrong tool, lost memory, unsafe action, context overflow.
Changeability	A pluggable part you can swap.	The durable architecture that outlives any single model.

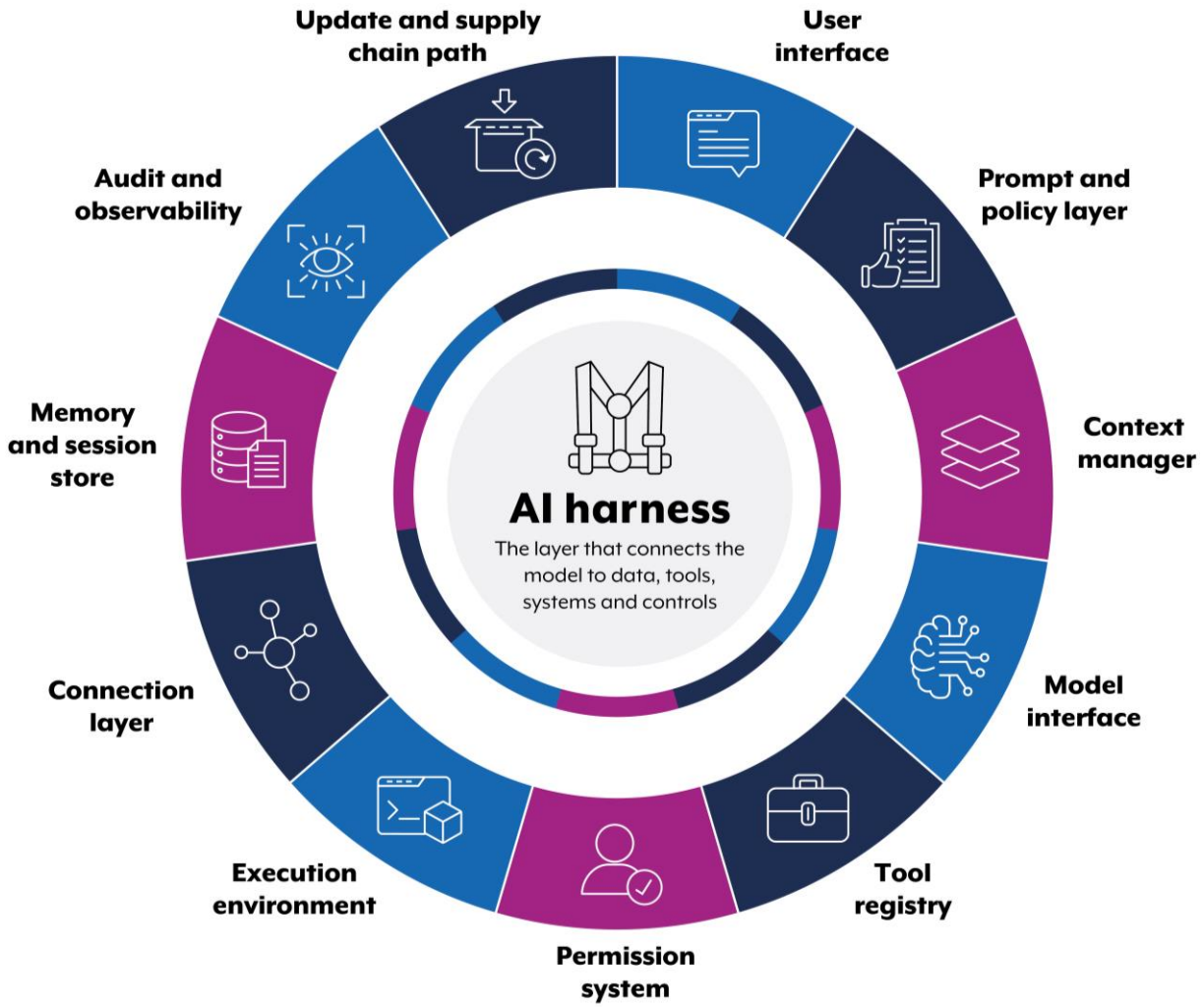
This distinction has two practical consequences.

First, the LLM is replaceable. A single harness can operate with multiple LLMs, and a well-designed harness will outlast successive model upgrades. For most organisations, the harness is the longer-term investment.

Second, many apparent agent failures result from the interaction between the model and the harness rather than the model alone. An LLM may select an inappropriate tool, use a tool incorrectly, lose context, exceed practical context limits, or propose an unsafe action. The harness determines which tools are available, enforces permissions and guardrails, and executes approved actions. For example, a model may propose a file deletion command, but whether that action is executed depends on the harness and its controls. When an agentic system behaves unexpectedly, organisations should investigate both the model's decision-making and the harness configuration rather than relying solely on prompt adjustments.

Harness components

The components of a harness, and the security relevance of each, are described in more detail below.



Component	Description
User interface	Determines the visibility users have of agent activity and the controls available for approval or intervention.
Prompt and policy layer	Stores system instructions, business rules and organisational policy, that help guide agent behaviour.
Context manager	Selects the information provided to the model and can expose sensitive information or omit important context.
Model interface	Provides the interface between the harness and organisation-approved LLM services, sending prompts, context and model parameters and receiving generated responses.

Component	Description
Tool registry	Defines available tools and their functions. Poorly defined tools may increase operational risk.
Permission system	Controls approvals and permissions for sensitive actions.
Execution environment	Defines where actions execute and can enforce bounds on potential actions.
Connector layer	Connects the harness to external data sources, services and enterprise systems, including RAG repositories and MCP-enabled tools.
Memory and session store	Retains information between sessions and may preserve sensitive, stale or manipulated content.
Audit and observability	Records activity for monitoring, review and incident response.
Update and supply chain path	Introduces new functionality and potential supply-chain risk.

These components are the organisation's configuration surface: most of the practical controls over an agentic AI system such as, which tools are exposed, what is permitted, which approvals are required and which connectors are attached, are settings within this list.

Security risks and mitigations

The harness both contributes to the attack surface of an agentic AI system and provides the means to defend against it. Every external tool, data source, connector and memory store the harness adds is a new path into the system. At the same time, the harness is where security controls must be applied, because the most significant weakness of agentic AI cannot be fixed in the model itself.

Broader security considerations

The harness increases both the capability and attack surface of an agentic AI system. By connecting LLMs to organisational data, tools, services and operational environments, it introduces additional trust relationships, dependencies and integration points that may be exploited by a malicious actor if not appropriately secured.

Agentic AI systems are inherently complex. Decisions and actions may depend on interactions between models, harnesses, tools, data sources and external services. This complexity can obscure the origin of failures and increase the likelihood of cascading impacts across interconnected systems.

Organisations should manage agentic AI security within established cyber security frameworks and apply existing security practices, including secure-by-design principles, defence in depth, identity and access management, monitoring and incident response. The harness is a key location where these controls can be implemented and enforced.

Agentic AI security risks

LLMs are provided instructions and information together as context when queried. As a result, a model may have difficulty distinguishing authorised instructions from information contained within external content, creating opportunities for prompt injection and other manipulation techniques. Content processed by an LLM, including web pages, documents, emails and code comments, may be interpreted as an instruction. This weakness underpins prompt injection attacks, for which no fully reliable technical mitigation currently exists. Because the weakness is inherent to how LLM's process context, mitigations must instead be applied in the harness, by controlling what an agent can access and what actions it is permitted to perform.

ASD's Careful adoption of agentic AI services categorises the security risks of agentic AI systems into five groups:

- **privilege risks**: agents granted excessive access or permissions, allowing a single compromise to have far-reaching consequences
- **design and configuration risks:** insecure architecture, integration or setup that creates weaknesses before the system is operational
- **behavioural risks:** agents pursuing goals in unintended ways, or being manipulated by malicious actors through techniques such as prompt injection and data poisoning
- **structural risks:** failures cascading across interconnected components and multi-step workflows, obscuring where a failure originated
- **accountability risks:** the complexity and opacity of agentic systems making it difficult to trace decisions, audit actions or assign responsibility.

For security purposes, a multi-agent system should be treated as a single agent, because a compromise in one component can propagate through shared context and trust relationships.

ASD's Careful adoption of agentic AI services recommends applying defence-in-depth controls across the agentic AI system lifecycle. Key practices include least-privilege access, strong identity and access management, controlled use of tools and external data sources, validation of agent outputs, human oversight of high-impact actions, continuous monitoring, supply-chain assurance and phased deployment of agentic capabilities. Organisations should implement these controls through the harness and surrounding infrastructure wherever possible, rather than relying solely on model behaviour or prompt instructions.

Many of these controls are implemented in, or enforced by, the harness.

Good practice

A harness can improve an organisation's cyber security posture by making security controls repeatable and enforceable, integrating security testing into workflows, limiting the effect of unsafe agent behaviour, and producing evidence that supports monitoring, investigation and continuous improvement.

- Use an organisation-controlled and hardened environment containing only the tools and permissions required for the task. Restrict access to production systems, sensitive data and network services.
- Configure the harness to apply organisational secure coding standards, prohibit unsafe actions and validate tool parameters and outputs. Require human approval for sensitive or high-impact actions.
- Verify outputs before putting them into operational usage. An agent's output and actions should be reviewed, tested and formally accepted before it is utilised in an operational setting, not treated as correct because it reads confidently. The harness may draft and act, but accountability for the result remains with a person.
- Implementing cost controls on APIs can help with cost monitoring as well security. Agentic sessions cost substantially more than a single exchange with a chatbot, whether billed as tokens by a hosted provider or consumed as computing resources on self hosted infrastructure. Monitoring this consumption can identify unexpected activity that may indicate misuse, compromise, runaway agent loops or denial-of-wallet attacks.
- Record prompts, responses, tool invocations, approvals, actions, security events and configuration changes. Protect, retain, review and independently monitor logs to support auditability, security monitoring, incident response, investigations and accountability
- Select a trusted model appropriate to the task and assess its provenance, limitations, security characteristics and behaviour. Do not rely on model safety controls instead of harness-enforced controls.
- Keep context focused and relevant, only provide information and access required for the task at hand. Operate the harness with least privilege. When managing AI workflows retain only information that remains relevant to the task. Where context must be reduced, delete stale content rather than summarising it, as summarisation rewrites the record and risks introducing new errors. Save durable facts and decisions to external memory or files, and replace a long, stale session with a fresh one seeded with a concise summary.
- Isolate exploration in sub-agents (a separate agent instance, the harness creates for a bounded part of a task) and promote their use. Applying least-privilege principles, organisations can configure subagents with only the tools, permissions and context required for specific task. This can limit exposure to untrusted content while keeping the primary agents session's context small and reducing security risk, operational cost and context degradation.

- Maintain a persistent mandatory rules file for the harness at an organisational level. Recording standards, prohibited areas, and build and test commands in a file the harness reads each session avoids rediscovering the same information every time, and makes the harness's behaviour more consistent and easier to review.

Governance questions for executives

Directors and senior executives do not need to understand every implementation detail of an agentic AI system. However, they should seek assurance that the harness, its integrations and its governance controls have been designed and implemented appropriately.

- What business outcome is this agentic AI system intended to achieve, and why is an agentic approach required?
- What data, systems and tools can the agent access, and are least-privilege principles being applied?
- What actions can the agent perform, and which actions require human approval?
- How are prompt injection, data poisoning and other AI-specific attacks being mitigated?
- Can all significant decisions, tool invocations and actions be monitored and audited?
- How will we know whether the system is delivering value, operating safely and remaining within acceptable risk tolerances?
- If the harness were compromised, misconfigured or manipulated, what is the worst outcome and what controls would prevent or limit that outcome?

Conclusion

The conversation around agentic AI often focuses on the LLM, but organisations should focus equally on the harness. The harness determines how an LLM interacts with tools, data and workflows, and it plays a major role in both capability and risk.

As models continue to change, the harness is likely to become the more enduring organisational capability. Well-designed harnesses can provide consistency, governance and security controls across changing generations of AI technology. Poorly designed harnesses can undermine even the most capable LLMs.

Start with low risk use cases, implement appropriate controls, maintain human oversight and expand deployment only as controls mature.

Organisations should apply the security practices described in *ASD's Careful adoption of agentic AI* services and implement them through the harness and surrounding ecosystem.

Additional reading

National Cyber Security Centre - UK

Prompt injection is not SQL injection (it may be worse)

Disclaimer

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

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

Copyright

Use of the Coat of Arms

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

Australian Government
Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC Australian Cyber Security Centre