CYBERGLASS SECURITY STANDARDS

AI Agent Security, Identity and Data Protection Standard

Mandatory controls at [Company Name] for AI agents: their identities and access, tools and connectors, prompt injection, data leakage, human approval and monitoring.

Organization	[Company Name]
Document ID	STD-04
Document owner	[Name, title]
Approved by	[Name, title]
Version	1.0 [draft / approved]
Effective date	[YYYY-MM-DD]
Next review	[YYYY-MM-DD] (at least every 12 months)
Classification	Internal

PREVIEW

About this template (delete this box before publishing)

This document is a template to help [Company Name] write its own policy. It is a guideline based on common industry practice and publicly available frameworks, not a finished policy, legal advice or a guarantee of compliance. Every organization is different: review each requirement, remove what does not apply, add what is missing, and set values your organization can actually meet. Have the final version reviewed by people who know your systems, obligations and risks.

- Use with the AI Governance Kit (AI-01 to AI-06) if you have it; this Standard also works on its own.
- Apply it to every agent before go-live; start with agents that can send, pay, delete or reach Restricted data.
- Give the worked example in Appendix B to developers as a model.
- Replace every highlighted **[placeholder]** with your organization's details, then remove the highlighting.
- Grey **Guidance** boxes explain what auditors, insurers and clients look for. Read them, then delete them before publishing.

Preview pages · full editable Word file at cyberglass.ai/shop

1. Purpose

This Standard sets mandatory requirements for building, deploying and operating AI agents: AI systems that reason over tasks and act through tools, data connectors and other systems on delegated authority. It covers agent identities and access, tool and connector hardening, protection against prompt injection, prevention of data leakage, human approval for high-impact actions, and monitoring. It applies the principles of the AI Acceptable Use Policy (AI-02) and the AI Risk and Impact Assessment Method (AI-03) at a technical level.

2. Objectives

The objectives of this Standard are to:

a) give every agent its own identity and only the access it needs for the task at hand, for as long as the task lasts
b) make sure a manipulated agent cannot reach data or take actions beyond its approved scope
c) keep untrusted content (web pages, emails, documents, tool results) from being treated as instructions
d) stop sensitive data leaving through agent outputs or actions
e) keep a human in control of high-impact actions and able to stop any agent at once
f) record what each agent did, on whose behalf and under what authority

3. Scope

This Standard applies to:

a) agents built in-house, agent features in purchased software where configuration is under [Company Name]'s control, and agents provided by vendors that act on [Company Name] systems or data
b) tools, plug-ins, data connectors and tool servers (including Model Context Protocol servers) used by agents
c) multi-agent workflows in which agents call or instruct other agents
d) all environments where agents can reach [Company Name] or client data

4. Agent Identity and Access

> **CONTROL ID** **STD-04-4** · requirements STD-04-4(a)–(g)
> **Maps to NIST CSF 2.0:** PR.AA-01, PR.AA-02, PR.AA-05
> **Maps to ISO/IEC 27001:2022:** A.5.16, A.5.18, A.8.2
> **Maps to NIST SP 800-53 Rev. 5:** IA-9, AC-6, IA-5
> **Maps to NIST AI RMF 1.0:** GOVERN 1, MAP 4, MANAGE 2

Every agent shall:

a) be registered in an agent register (the Agent Register sheet of the AI Governance Workbook, or your own) with a human owner, purpose, risk tier (AI-03), tools, data classifications and approval gates
b) have its own non-human identity; agents shall not use a person's credentials or a shared service account
c) receive short-lived access tokens scoped to the task, the user on whose behalf it acts and the minimum data needed, issued by the identity provider or a token service at the start of each task (for example OAuth token exchange or workload identity), with a lifetime of at most [60] minutes
d) have no standing privileges: access ends when the task ends, and permanent API keys are not used where a short-lived token is possible
e) never be able to grant itself or another agent wider access than it was given

Preview pages · full editable Word file at cyberglass.ai/shop

f) authenticate to tools, APIs and other agents with credentials bound to the agent's workload (for example mutual TLS certificates or signed, audience-restricted tokens), so a stolen token cannot be replayed from another host or used against another service

g) keep secrets in a secrets vault; no keys, passwords or tokens in prompts, configuration files, code repositories or agent memory

Agent risk tier (AI-03)	Identity and token rules	Access review
Low	Own identity; read-only scopes; token ≤ [8] h	Every [6] months
Medium	Task-scoped token ≤ [60] min; on-behalf-of user scopes	Quarterly
High	Task-scoped token ≤ [15] min; approval gates (section 8); separate identity per deployment	Monthly, plus after any change

SMB · Small organization (about 20–250 staff)	ENTERPRISE · Large enterprise or regulated operator
Use the AI platform's built-in agent identities and scopes; no agent gets send, pay or delete rights.	Central agent identity service issuing task-scoped tokens, policy engine before each tool call, and quarterly agent access certification.

AUDITOR RED FLAG *What auditors, customers' security reviewers and AI management system auditors check here (delete before publishing)*

What they look for

- Each agent with its own identity, owner and least-privilege scopes

Evidence that usually satisfies them

- Agent register; identity provider list of agent identities with scopes and token lifetimes

Red flags that lead to findings

- Agents running on a developer's personal token
- Shared service account across agents
- API keys in code repositories

5. Delegation and Accountability

CONTROL ID **STD-04-5** · requirements STD-04-5(a)–(g)
Maps to NIST CSF 2.0: PR.PS-04, DE.CM-09
Maps to ISO/IEC 27001:2022: A.8.15
Maps to NIST SP 800-53 Rev. 5: AU-2, AU-10
Maps to NIST AI RMF 1.0: MEASURE 3

For every agent action that reads or changes data or calls an external service, the agent platform shall record:

a) the agent identity and version, and the human user or system on whose behalf it acted

b) who delegated the task and the scope delegated, including each hop when one agent calls another

c) the tool or connector called, the parameters (with sensitive values masked) and the result status

d) any policy decision taken (allowed, blocked, escalated) and any human approval

e) the model and prompt template versions in use

Preview pages · full editable Word file at cyberglass.ai/shop

f) references to the sources the agent retrieved or read for the action (document IDs, URLs or chunk identifiers), so investigators can find content that may have influenced it

g) reasoning summaries where the platform provides them; these help investigation but may be incomplete and are not proof of why the model acted

Logs are retained for at least [12] months, protected from change by the agent, and reviewed under section 10.

6. Tool, Connector and Tool-Server Hardening

CONTROL ID **STD-04-6** · requirements STD-04-6(a)–(h)
Maps to NIST CSF 2.0: PR.PS-01, PR.PS-05, PR.AA-05
Maps to ISO/IEC 27001:2022: A.8.9, A.8.19, A.5.23
Maps to NIST SP 800-53 Rev. 5: CM-7, SC-39, SA-9
Maps to NIST AI RMF 1.0: MANAGE 2, GOVERN 6

Tools and connectors available to agents shall:

a) be on an approved allowlist per agent; wildcard permissions are not allowed

b) be read-only by default; write, send, delete and execute rights are granted only where the use case requires them and the risk assessment approves

c) limit file access to named paths (for example `/app/reports/*`) and deny secrets and configuration patterns (for example `*.env`, `*.key`, `*.pem`, `*secret*`, `*credential*`)

d) limit database access to named schemas, views or stored procedures, with row and column limits, and deny unrestricted queries and schema changes

e) not provide general shell or code execution, unless the code runs in an isolated, short-lived sandbox (for example a micro-virtual machine or container) with no network access except approved destinations, no stored secrets and resource limits

f) validate tool inputs against an expected format before execution and validate tool outputs before they are passed to other systems

g) be verified before installation and at each update: signature or hash of the tool-server or plug-in package checked against the approved version, and manifests describing each tool's permissions reviewed for changes

h) come only from approved sources: third-party tool servers, plug-ins and connectors are assessed under AI-04, pinned to reviewed versions and monitored for changes

AUDITOR RED FLAG *What independent assessors and auditors check here (delete before publishing)*

What they look for

- Tool allowlists and sandboxing

Evidence that usually satisfies them

- Tool configuration per agent; sandbox settings

Red flags that lead to findings

- Wildcard file or database access
- Shell access in production
- Unreviewed third-party tool servers

PREVIEW

Preview pages · full editable Word file at cyberglass.ai/shop