OLYMPUS CYBER

AI AGENT

SECURITY CONTROL LIBRARY

28 Controls. 8 Domains. Mapped to NIST CSF, ISO 27001, CSA CCM, PCI DSS, and CIS.

A public framework for securing autonomous agents before production.

Released	August 2026
Version	2.0
Controls	28 across 8 domains
Audience	Security leaders and practitioners
Classification	PUBLIC

FREE PUBLIC RELEASE
Use it. Adapt it. Import it into your control library.

Executive Cyber Resilience for the AI Era

olympus-cyber.com

EXECUTIVE SUMMARY

Organizations are deploying autonomous agents into production faster than they are securing them. The agent gets access to systems, credentials, and decision authority that no contractor would receive without a background check, and it gets that access because the deployment was treated as a productivity project rather than a change to the control environment.

This library exists to close that gap. It defines 28 controls across 8 domains, each mapped to NIST CSF 2.0, ISO 27001:2022, CSA Cloud Controls Matrix v4, PCI DSS 4.0, and CIS Controls v8.1. It is written to be imported directly into an existing GRC platform or control library, not read once and filed.

What this library assumes

- The agent will be compromised or will malfunction at some point. Design for containment, not for prevention alone.
- Prompt injection cannot be fully solved at the model layer. Authorization has to be the backstop.
- An agent decision you cannot reconstruct is an agent decision you cannot defend.
- Agent scope expands quietly. Governance has to be periodic, not one-time.

Who should use it

Security leaders use it to establish deployment gates and report agent risk posture to the board. Practitioners use it as an architecture review and adversarial test checklist. Audit and compliance teams use the framework mappings to fold agent controls into existing SOC 2, ISO 27001, PCI DSS, and CMMC programs without creating a parallel control set.

License

HOW TO USE THIS LIBRARY

Controls are assigned a tier so you can sequence adoption rather than attempting all 28 at once.

Tier	Meaning	When
Tier 1	Foundational. The agent should not reach production without these.	Before first production deployment
Tier 2	Core. Required for sustained production operation at any meaningful scope.	Within 90 days of deployment
Tier 3	Maturity. Detection and analytics that separate a managed agent from a monitored one.	As the program matures

Import guidance

The accompanying CSV and JSON files carry the same 28 controls in a structured format. Map the columns to your platform as follows.

Library field	Typical GRC target field
control_id	Control ID or unique reference
control_name	Control title
domain	Control family, domain, or category
tier	Priority, phase, or implementation wave
control_statement	Control description or requirement
implementation_guidance	Implementation notes or testing procedure
risk_if_absent	Risk statement or rationale
nist_csf_2_0 / iso_27001_2022 / csa_ccm_v4 / pci_dss_4_0 / cis_v8_1	Framework crosswalk or mapped authority document

QUICKSTART: THREE GATES

Before reading further, answer these three questions about any agent you have already deployed. If any answer is no, you have a production gap today.

Gate	Question	Control
1	Can you disable this agent in under 30 seconds, without an engineer?	AI-ASC-25
2	Can you list every system, dataset, and tool the agent can reach right now?	AI-ASC-08
3	Can you reconstruct why the agent made a specific decision last Tuesday?	AI-ASC-16

Most organizations fail at least one of these on their first pass. That is the point of the exercise.

CONTROL INDEX

All 28 controls at a glance.

ID	Control	Domain	Tier
AI-ASC-01	Network Segmentation for Agent Runtime	Isolation and Boundary Protection	Tier 1
AI-ASC-02	Credential Isolation and Vaulting	Isolation and Boundary Protection	Tier 1
AI-ASC-03	API Gateway Mediation	Isolation and Boundary Protection	Tier 1
AI-ASC-04	Resource Quota Enforcement	Isolation and Boundary Protection	Tier 2
AI-ASC-05	Process and Container Isolation	Isolation and Boundary Protection	Tier 2
AI-ASC-06	Least Privilege Agent Identity	Authentication and Authorization	Tier 1
AI-ASC-07	Non-Human Identity Registration	Authentication and Authorization	Tier 1
AI-ASC-08	Documented Permission Inventory	Authentication and Authorization	Tier 2
AI-ASC-09	Credential Rotation and Expiry	Authentication and Authorization	Tier 2
AI-ASC-10	Schema Validation on All Inputs	Input Validation and Data Integrity	Tier 1
AI-ASC-11	Data Source Integrity Verification	Input Validation and Data Integrity	Tier 2
AI-ASC-12	Prompt Injection Prevention	Input Validation and Data Integrity	Tier 1
AI-ASC-13	Input Anomaly Detection	Input Validation and Data Integrity	Tier 3
AI-ASC-14	Comprehensive Decision Logging	Decision Logging and Audit	Tier 1
AI-ASC-15	Immutable Audit Trail	Decision Logging and Audit	Tier 2
AI-ASC-16	Forensic Reconstruction Capability	Decision Logging and Audit	Tier 2
AI-ASC-17	Confidence and Uncertainty Capture	Decision Logging and Audit	Tier 3
AI-ASC-18	Graceful Degradation on Failure	Failure Handling and Escalation	Tier 1
AI-ASC-19	Defined Escalation Thresholds	Failure Handling and Escalation	Tier 1
AI-ASC-20	Human Response Service Level	Failure Handling and Escalation	Tier 2
AI-ASC-21	Rollback and State Recovery	Failure Handling and Escalation	Tier 2
AI-ASC-22	Behavioral Baseline and Analytics	Monitoring and Detection	Tier 2
AI-ASC-23	Output Validation and Hallucination Detection	Monitoring and Detection	Tier 2
AI-ASC-24	Agent Health Heartbeat	Monitoring and Detection	Tier 3
AI-ASC-25	Instant Kill Switch	Incident Response and Kill Switch	Tier 1
AI-ASC-26	Agent Incident Response Playbook	Incident Response and Kill Switch	Tier 2
AI-ASC-27	Model, Prompt, and Tool Version Control	Governance and Change Management	Tier 2
AI-ASC-28	Periodic Agent Risk Review	Governance and Change Management	Tier 2

DOMAIN 1: ISOLATION AND BOUNDARY PROTECTION

Contain the agent. If it is compromised or misbehaves, the damage stops at a boundary you defined in advance.

5 controls: AI-ASC-01, AI-ASC-02, AI-ASC-03, AI-ASC-04, AI-ASC-05

AI-ASC-01 Network Segmentation for Agent Runtime

Tier	Tier 1
Control statement	The agent executes inside a dedicated network segment with explicit allow-list egress. It cannot reach production systems, identity infrastructure, or data stores that are not required by its documented function.
Implementation guidance	Place the agent runtime in its own VLAN, VPC subnet, or namespace. Default-deny all egress and add explicit allow rules per destination and port. Route all outbound traffic through an inspected proxy. Review the allow-list whenever the agent scope changes.
Risk if absent	An agent without network containment becomes a lateral movement platform. A single prompt injection or dependency compromise gives an attacker the agent's full network reach.
Framework mapping	**NIST CSF 2.0:** PR.IR-01, PR.AA-05 **ISO 27001:2022:** 8.20, 8.22 **CSA CCM v4:** IVS-03, IVS-07 **PCI DSS 4.0:** 1.2, 1.3, 1.4 **CIS Controls v8.1:** 12.2, 12.4, 4.4

AI-ASC-02 Credential Isolation and Vaulting

Tier	Tier 1
Control statement	Agent credentials are stored in a secrets manager, injected at runtime, and never present in prompts, source code, configuration files, model context, or logs.
Implementation guidance	Use a managed vault (HashiCorp Vault, AWS Secrets Manager, Azure Key Vault). Inject secrets as short-lived environment values or via workload identity federation. Scan repositories, prompt templates, and log pipelines for credential leakage. Deny the agent read access to its own secret store.
Risk if absent	Credentials inside model context can be extracted through prompt injection or leaked into logs and vendor telemetry. Recovery requires rotating every credential the agent has ever held.
Framework mapping	**NIST CSF 2.0:** PR.AA-01, PR.DS-01 **ISO 27001:2022:** 5.17, 8.24 **CSA CCM v4:** IAM-06, CEK-03 **PCI DSS 4.0:** 3.5, 8.3.1 **CIS Controls v8.1:** 3.11, 5.2

AI-ASC-03 API Gateway Mediation

Tier	Tier 1
Control statement	All agent calls to external systems pass through a gateway that enforces authentication, authorization, rate limits, and request logging. The agent holds no direct connections to backend systems.
Implementation	Front every agent-reachable API with a gateway. Enforce per-endpoint authorization at the

guidance	gateway rather than trusting agent-side logic. Log full request and response metadata. Fail closed when the gateway is unavailable.
Risk if absent	Direct backend access removes the enforcement point. You lose the ability to throttle, audit, or revoke agent behavior without redeploying the agent itself.
Framework mapping	**NIST CSF 2.0:** PR.AA-05, PR.PS-01 **ISO 27001:2022:** 8.21, 8.26 **CSA CCM v4:** AIS-02, IVS-06 **PCI DSS 4.0:** 1.3, 6.4.2 **CIS Controls v8.1:** 12.2, 16.10

AI-ASC-04 Resource Quota Enforcement

Tier	Tier 2
Control statement	Hard ceilings are enforced on agent token consumption, API call volume, execution duration, concurrent instances, and compute spend. Breaching a ceiling halts the agent rather than degrading it.
Implementation guidance	Set quotas at the gateway and the platform layer, not in agent logic. Define per-hour and per-day ceilings for each. Alert at 70 percent and halt at 100 percent. Require human approval to raise a ceiling.
Risk if absent	An agent in a reasoning loop can generate six-figure cost events and denial of service against your own systems in under an hour. Without a hard ceiling there is nothing between a logic error and the invoice.
Framework mapping	**NIST CSF 2.0:** PR.IR-04, DE.CM-09 **ISO 27001:2022:** 8.6 **CSA CCM v4:** IVS-04, IPY-03 **PCI DSS 4.0:** 12.10.5 **CIS Controls v8.1:** 12.1, 13.1

AI-ASC-05 Process and Container Isolation

Tier	Tier 2
Control statement	The agent runs in an ephemeral, non-privileged container or sandbox with a read-only filesystem, no host mounts, and no ability to spawn arbitrary processes outside its execution boundary.
Implementation guidance	Run as a non-root user with all Linux capabilities dropped. Mount the root filesystem read-only with a scoped writable temp volume. Disable host network and host PID namespaces. Rebuild the container from a signed image on every run so state does not persist between sessions.
Risk if absent	A persistent, privileged runtime lets an attacker establish durable footholds. Tool-execution agents that can write to disk and spawn processes are functionally remote code execution as a service.
Framework mapping	**NIST CSF 2.0:** PR.PS-01, PR.PS-05 **ISO 27001:2022:** 8.31, 8.19 **CSA CCM v4:** IVS-09, CCC-04 **PCI DSS 4.0:** 2.2, 6.4.1 **CIS Controls v8.1:** 4.1, 4.6, 2.7

DOMAIN 2: AUTHENTICATION AND AUTHORIZATION

The agent is a named identity with a defined and reviewable set of rights. It is not a shared account and not an administrator.

4 controls: AI-ASC-06, AI-ASC-07, AI-ASC-08, AI-ASC-09

AI-ASC-06 Least Privilege Agent Identity

Tier	Tier 1
Control statement	The agent holds a dedicated service identity scoped to the minimum permissions required for its documented function. It does not inherit or reuse human, administrator, or shared application credentials.
Implementation guidance	Create a unique service principal or IAM role per agent, per environment. Start from zero permissions and add only what a documented workflow requires. Prohibit wildcard scopes and directory-wide or tenant-wide roles. Re-derive permissions from scratch whenever function changes rather than appending.
Risk if absent	Agents provisioned with administrator rights for convenience convert every reasoning error into a privileged action. Shared credentials also make attribution impossible during an investigation.
Framework mapping	**NIST CSF 2.0:** PR.AA-05, PR.AA-01 **ISO 27001:2022:** 5.15, 8.2 **CSA CCM v4:** IAM-05, IAM-16 **PCI DSS 4.0:** 7.2, 7.3 **CIS Controls v8.1:** 6.8, 5.4

AI-ASC-07 Non-Human Identity Registration

Tier	Tier 1
Control statement	Every agent identity is registered in an inventory with a named human owner, business purpose, permission scope, data classification reached, and decommission date.
Implementation guidance	Extend the existing asset or identity inventory rather than building a separate list. Require the owner field to be a named individual, not a team alias. Block production deployment of any agent identity that is not registered. Reconcile the inventory against actual directory objects quarterly.
Risk if absent	Unregistered agent identities are the shadow IT of the AI era. You cannot revoke, review, or investigate an identity you do not know exists, and orphaned agents outlive the projects that created them.
Framework mapping	**NIST CSF 2.0:** ID.AM-01, GV.RR-02 **ISO 27001:2022:** 5.9, 5.16 **CSA CCM v4:** IAM-01, DCS-06 **PCI DSS 4.0:** 8.1, 12.5.1 **CIS Controls v8.1:** 1.1, 5.1

AI-ASC-08 Documented Permission Inventory

Tier	Tier 2
Control statement	A current, human-readable list exists of every API, dataset, tool, and system the agent can reach, with the specific action verbs permitted on each.
Implementation	Generate the list from live configuration rather than maintaining it by hand. State read,

guidance	write, delete, and execute rights separately per resource. Publish it to the control owner and to audit. Treat any drift between the document and live configuration as a control failure.
Risk if absent	If you cannot enumerate agent access on demand, you cannot scope an incident. Investigations stall at the question of what the agent could have touched, and regulators read that gap as a lack of governance.
Framework mapping	**NIST CSF 2.0:** ID.AM-02, GV.OC-04 **ISO 27001:2022:** 5.9, 8.9 **CSA CCM v4:** IAM-02, GRC-05 **PCI DSS 4.0:** 7.2.4, 12.5.2 **CIS Controls v8.1:** 1.1, 6.1

AI-ASC-09 Credential Rotation and Expiry

Tier	Tier 2
Control statement	Agent credentials are short-lived and rotate automatically. Long-lived static keys and non-expiring tokens are prohibited.
Implementation guidance	Prefer workload identity federation or OIDC token exchange over static keys. Where static credentials are unavoidable, cap lifetime at 90 days and automate rotation. Alert on any agent credential older than policy. Revoke immediately on agent decommission.
Risk if absent	A leaked static agent key is a permanent backdoor. Agent credentials are especially exposed because they pass through prompts, logs, vendor telemetry, and third-party tool integrations.
Framework mapping	**NIST CSF 2.0:** PR.AA-01, PR.AA-02 **ISO 27001:2022:** 5.17, 8.5 **CSA CCM v4:** IAM-14, CEK-08 **PCI DSS 4.0:** 8.3.9, 8.6.3 **CIS Controls v8.1:** 5.2, 5.3

DOMAIN 3: INPUT VALIDATION AND DATA INTEGRITY

Everything reaching the agent is untrusted input, including data from your own systems. Treat instructions embedded in data as hostile by default.

4 controls: AI-ASC-10, AI-ASC-11, AI-ASC-12, AI-ASC-13

AI-ASC-10 Schema Validation on All Inputs

Tier	Tier 1
Control statement	Structured inputs are validated against an explicit schema before the agent processes them. Inputs failing validation are rejected and logged rather than passed through.
Implementation guidance	Define schemas for every input channel including tool outputs and retrieved documents. Enforce type, length, encoding, and range. Reject on failure, do not coerce or truncate silently. Log rejections with source attribution so you can detect probing.
Risk if absent	Malformed and oversized inputs drive unpredictable model behavior and are the delivery mechanism for injection payloads. Silent coercion hides the attack from your logs.
Framework mapping	**NIST CSF 2.0:** PR.DS-02, PR.PS-06 **ISO 27001:2022:** 8.26, 8.28 **CSA CCM v4:** AIS-04, AIS-07 **PCI DSS 4.0:** 6.2.4, 6.4.1 **CIS Controls v8.1:** 16.10, 16.11

AI-ASC-11 Data Source Integrity Verification

Tier	Tier 2
Control statement	Every data source the agent consumes is explicitly approved, authenticated, and integrity-checked. The agent cannot fetch from arbitrary or user-supplied locations.
Implementation guidance	Maintain an allow-list of retrieval sources. Require TLS with certificate validation on every fetch. Verify signatures or checksums where the source supports it. Prohibit agent-initiated fetches to URLs supplied at runtime by users or by other content.
Risk if absent	Poisoned retrieval sources let an attacker steer agent decisions without ever touching your network. Retrieval-augmented agents inherit the trustworthiness of the weakest document in the index.
Framework mapping	**NIST CSF 2.0:** PR.DS-01, ID.RA-09 **ISO 27001:2022:** 5.23, 8.24 **CSA CCM v4:** DSP-03, STA-06 **PCI DSS 4.0:** 6.4.3, 12.8 **CIS Controls v8.1:** 3.10, 15.4

AI-ASC-12 Prompt Injection Prevention

Tier	Tier 1
Control statement	System instructions are structurally separated from untrusted content, and instruction-like text appearing inside retrieved or user-supplied data is neutralized rather than executed.
Implementation guidance	Use the platform's structured system and role separation rather than string concatenation. Wrap untrusted content in explicit delimiters and instruct the model to treat it as data. Filter known injection patterns at ingest. Enforce authorization on the action itself so a successful injection still cannot exceed the agent's rights.

Risk if absent	Prompt injection is the primary attack path against agentic systems and it does not require network access. Defense that relies only on model instruction-following will fail. Authorization must be the backstop.
Framework mapping	**NIST CSF 2.0:** PR.PS-06, DE.CM-09 **ISO 27001:2022:** 8.26, 8.28 **CSA CCM v4:** AIS-04, TVM-02 **PCI DSS 4.0:** 6.2.4 **CIS Controls v8.1:** 16.11, 16.12

AI-ASC-13 Input Anomaly Detection

Tier	Tier 3
Control statement	Input volume, size, source distribution, and content patterns are baselined, and deviations trigger alerts to a monitored queue.
Implementation guidance	Baseline over at least 30 days of normal operation. Alert on sudden volume shifts, unusual source addresses, abnormal payload sizes, and spikes in rejected inputs. Route alerts to the same queue that receives security telemetry, not to an unmonitored inbox.
Risk if absent	Injection campaigns and data poisoning attempts show up as input anomalies well before they show up as bad decisions. Without baselining, the first signal is the incident.
Framework mapping	**NIST CSF 2.0:** DE.AE-02, DE.CM-01 **ISO 27001:2022:** 8.16 **CSA CCM v4:** LOG-13, TVM-03 **PCI DSS 4.0:** 10.4, 11.5 **CIS Controls v8.1:** 13.1, 8.11

DOMAIN 4: DECISION LOGGING AND AUDIT

Every agent decision must be reconstructable after the fact. If you cannot explain what the agent did and why, you cannot defend it to a board, a regulator, or opposing counsel.

4 controls: AI-ASC-14, AI-ASC-15, AI-ASC-16, AI-ASC-17

AI-ASC-14 Comprehensive Decision Logging

Tier	Tier 1
Control statement	Every agent decision is logged with timestamp, input received, tools invoked, reasoning or rationale captured, action taken, and outcome returned.
Implementation guidance	Log at the orchestration layer so coverage does not depend on the model. Capture tool call arguments and results, not just the final answer. Assign a correlation identifier per session and per decision chain. Redact secrets and regulated data at write time rather than after.
Risk if absent	Logs that record only the final action leave the reasoning invisible. When an agent makes a costly decision, the missing middle is exactly what the investigation, the insurer, and the regulator will ask for.
Framework mapping	**NIST CSF 2.0:** DE.AE-03, PR.PS-04 **ISO 27001:2022:** 8.15 **CSA CCM v4:** LOG-03, LOG-05 **PCI DSS 4.0:** 10.2.1, 10.2.2 **CIS Controls v8.1:** 8.2, 8.5

AI-ASC-15 Immutable Audit Trail

Tier	Tier 2
Control statement	Agent logs are written to append-only, tamper-evident storage that neither the agent nor its service identity can modify or delete.
Implementation guidance	Ship logs off the agent host in near real time. Use write-once storage or object lock. Explicitly deny the agent identity any write or delete permission on the log store. Set retention to the longer of one year or your regulatory obligation.
Risk if absent	A compromised agent with write access to its own logs can erase the evidence of its compromise. Mutable logs also fail evidentiary standards, which undermines both claims and litigation positions.
Framework mapping	**NIST CSF 2.0:** PR.DS-01, DE.AE-06 **ISO 27001:2022:** 8.15, 5.28 **CSA CCM v4:** LOG-09, LOG-10 **PCI DSS 4.0:** 10.3.2, 10.5.1 **CIS Controls v8.1:** 8.3, 8.9

AI-ASC-16 Forensic Reconstruction Capability

Tier	Tier 2
Control statement	The organization can reconstruct the complete decision chain for any agent action within the retention window, and this capability is tested rather than assumed.
Implementation guidance	Test quarterly by picking a random action from the prior week and reconstructing it end to end. Record how long the reconstruction takes. Fix log gaps found during the test. Document the reconstruction procedure so it does not depend on one engineer.

Risk if absent	Logging that has never been exercised usually has gaps. Discovering them during a live incident costs days at the point where hours matter, and turns a contained event into a disclosure decision made without facts.
Framework mapping	**NIST CSF 2.0:** RS.AN-03, DE.AE-02 **ISO 27001:2022:** 5.28, 8.15 **CSA CCM v4:** SEF-06, LOG-08 **PCI DSS 4.0:** 10.4.1, 12.10.1 **CIS Controls v8.1:** 8.11, 17.4

AI-ASC-17 Confidence and Uncertainty Capture

Tier	Tier 3
Control statement	Where the agent produces a confidence signal, that value is recorded with the decision and used to drive escalation thresholds.
Implementation guidance	Persist confidence or uncertainty scores alongside each logged decision. Analyze the relationship between low confidence and downstream errors on a monthly basis. Tune escalation thresholds from that data rather than from intuition. Treat consistently low-confidence workflows as candidates for removal from automation.
Risk if absent	Without confidence data, every agent decision looks equally sound in the log. You lose the earliest available signal that the agent is operating outside its competence.
Framework mapping	**NIST CSF 2.0:** ID.RA-05, DE.AE-04 **ISO 27001:2022:** 8.16, 5.7 **CSA CCM v4:** LOG-05, GRC-03 **PCI DSS 4.0:** 12.3.1 **CIS Controls v8.1:** 8.11

DOMAIN 5: FAILURE HANDLING AND ESCALATION

The agent's failure mode must be to stop and ask, never to guess and proceed.

4 controls: AI-ASC-18, AI-ASC-19, AI-ASC-20, AI-ASC-21

AI-ASC-18 Graceful Degradation on Failure

Tier	Tier 1
Control statement	When a dependency fails, a tool errors, or the agent cannot complete a step, it halts, logs the failure, and escalates. It does not substitute assumptions and continue.
Implementation guidance	Define explicit failure handling for every tool call including timeouts and partial responses. Fail closed by default. Prohibit retry loops without a bounded attempt count and a backoff. Make the halted state visible in the operational dashboard.
Risk if absent	Agents that continue through failure produce confident output built on missing data. The output looks identical to a correct result, which is why the error is usually found downstream by a customer or an auditor.
Framework mapping	**NIST CSF 2.0:** RS.MA-02, PR.PS-04 **ISO 27001:2022:** 8.14, 8.16 **CSA CCM v4:** BCR-05, AIS-07 **PCI DSS 4.0:** 10.7.2, 12.10.5 **CIS Controls v8.1:** 17.3, 16.12

AI-ASC-19 Defined Escalation Thresholds

Tier	Tier 1
Control statement	Documented thresholds determine which agent decisions require human approval before execution, based on blast radius, data sensitivity, financial value, and reversibility.
Implementation guidance	Write the thresholds before deployment and have the business owner approve them. Require human authorization for anything irreversible, anything touching regulated data, and anything above a defined financial value. Enforce the threshold at the gateway, not in the prompt. Review thresholds after every escalation event.
Risk if absent	Undefined thresholds mean the agent's boundary is whatever the model decides in the moment. That is not a control, and it will not survive scrutiny after an adverse outcome.
Framework mapping	**NIST CSF 2.0:** GV.RR-01, PR.AA-05 **ISO 27001:2022:** 5.3, 5.15 **CSA CCM v4:** GRC-04, IAM-10 **PCI DSS 4.0:** 7.2.1, 12.5.2 **CIS Controls v8.1:** 6.8, 14.6

AI-ASC-20 Human Response Service Level

Tier	Tier 2
Control statement	Escalations route to a named, staffed queue with a defined response time. Unacknowledged escalations trigger a secondary path rather than expiring.
Implementation guidance	Define response targets by severity and publish them. Route to an on-call rotation, not to an individual. Auto-escalate to a secondary contact on breach. Track and report acknowledgment time monthly as a control metric.
Risk if absent	An escalation path nobody watches is functionally an unmonitored agent. Teams under pressure begin approving escalations without review, which converts the control into a

	rubber stamp.
Framework mapping	**NIST CSF 2.0:** RS.CO-02, GV.RR-02 **ISO 27001:2022:** 5.24, 5.26 **CSA CCM v4:** SEF-02, SEF-03 **PCI DSS 4.0:** 12.10.1, 12.10.3 **CIS Controls v8.1:** 17.2, 17.4

AI-ASC-21 Rollback and State Recovery

Tier	Tier 2
Control statement	Every environment the agent can modify has a tested restore path to a known good state, with a documented and measured recovery time.
Implementation guidance	Snapshot or version any state the agent can write. Test the restore quarterly and record actual recovery time. Ensure the restore path does not depend on the agent or on credentials the agent holds. Confirm backups are outside the agent's write scope.
Risk if absent	An agent operating at machine speed can propagate an error across thousands of records before anyone notices. Without a tested rollback, remediation becomes manual reconstruction measured in weeks.
Framework mapping	**NIST CSF 2.0:** RC.RP-01, RC.RP-03 **ISO 27001:2022:** 8.13, 5.29 **CSA CCM v4:** BCR-08, BCR-11 **PCI DSS 4.0:** 12.10.1 **CIS Controls v8.1:** 11.3, 11.5

DOMAIN 6: MONITORING AND DETECTION

Agent behavior drifts. Detection has to be continuous, because the agent that passed testing is not the agent running in month six.

3 controls: AI-ASC-22, AI-ASC-23, AI-ASC-24

AI-ASC-22 Behavioral Baseline and Analytics

Tier	Tier 2
Control statement	Normal agent behavior is baselined across decision volume, tool usage mix, error rate, escalation rate, and resource consumption. Deviations generate alerts.
Implementation guidance	Establish the baseline over at least 30 days of steady-state operation. Alert on statistically significant deviation, not on fixed thresholds alone. Feed agent telemetry into the SIEM alongside other security data. Re-baseline after any model, prompt, or scope change.
Risk if absent	Agent compromise and agent malfunction present identically in the early stage as behavioral drift. Without a baseline there is no deviation to detect, and the first indicator is business impact.
Framework mapping	**NIST CSF 2.0:** DE.CM-01, DE.AE-02 **ISO 27001:2022:** 8.16 **CSA CCM v4:** LOG-13, TVM-03 **PCI DSS 4.0:** 10.4.1, 10.6 **CIS Controls v8.1:** 8.11, 13.1

AI-ASC-23 Output Validation and Hallucination Detection

Tier	Tier 2
Control statement	Agent outputs are validated against schema, business rules, and where feasible an authoritative source before the output is acted upon.
Implementation guidance	Enforce output schema at the orchestration layer. Cross-check factual claims and identifiers against systems of record before execution. Flag outputs referencing entities, tickets, or accounts that do not exist. Sample and human-review a fixed percentage of outputs on an ongoing basis.
Risk if absent	A fabricated identifier or invented finding executed at machine speed becomes an operational and legal problem, not a model quirk. Downstream systems generally accept plausible input without question.
Framework mapping	**NIST CSF 2.0:** DE.CM-09, PR.DS-06 **ISO 27001:2022:** 8.26, 8.29 **CSA CCM v4:** AIS-07, DSP-16 **PCI DSS 4.0:** 6.2.4, 11.6 **CIS Controls v8.1:** 16.11, 16.13

AI-ASC-24 Agent Health Heartbeat

Tier	Tier 3
Control statement	The agent emits a periodic health signal. Loss of heartbeat generates an alert within a defined window.
Implementation guidance	Emit a heartbeat at an interval matched to the agent's risk level. Alert on missed beats to a monitored queue. Include queue depth, last successful action, and error count in the signal. Confirm the heartbeat itself is monitored, not merely emitted.

Risk if absent	A silently stopped agent creates a coverage gap nobody knows about. Work assumed to be handled simply is not, which is most dangerous when the agent performs a monitoring or triage function.
Framework mapping	**NIST CSF 2.0:** DE.CM-01, PR.PS-04 **ISO 27001:2022:** 8.16, 8.6 **CSA CCM v4:** LOG-08, IVS-04 **PCI DSS 4.0:** 10.7.1, 10.7.2 **CIS Controls v8.1:** 8.11, 13.1

DOMAIN 7: INCIDENT RESPONSE AND KILL SWITCH

You must be able to stop the agent immediately, and you must have decided in advance who does it and what happens next.

2 controls: AI-ASC-25, AI-ASC-26

AI-ASC-25 Instant Kill Switch

Tier	Tier 1
Control statement	A single human-executable action disables the agent completely within 30 seconds, revoking credentials and terminating in-flight operations. It requires no engineering support and no code deployment.
Implementation guidance	Implement disablement at the identity and gateway layer so it does not depend on the agent responding. Document the exact command or console path and store it where on-call can reach it. Grant kill authority to on-call, not only to the agent owner. Test the kill switch monthly and record the measured time.
Risk if absent	A kill switch that requires a code change or a specific engineer is not a kill switch. The interval between recognizing a runaway agent and stopping it is the entire blast radius of the event.
Framework mapping	**NIST CSF 2.0:** RS.MI-01, RS.MI-02 **ISO 27001:2022:** 5.26, 8.16 **CSA CCM v4:** SEF-05, IAM-14 **PCI DSS 4.0:** 12.10.1, 8.2.6 **CIS Controls v8.1:** 17.5, 5.3

AI-ASC-26 Agent Incident Response Playbook

Tier	Tier 2
Control statement	A written playbook covers agent-specific incidents including compromise, runaway behavior, data exposure through agent action, and injection-driven unauthorized activity. It has been exercised.
Implementation guidance	Extend the existing IR plan rather than writing a separate one. Define agent-specific triage questions, evidence sources, and containment steps. Name the decision authority for shutting down a revenue-generating agent. Run a tabletop exercise at least annually and after any material scope change.
Risk if absent	Generic IR plans do not answer agent-specific questions such as whether output produced during the compromise window is trustworthy. That question drives customer notification and it will be asked under time pressure.
Framework mapping	**NIST CSF 2.0:** RS.MA-01, ID.IM-02 **ISO 27001:2022:** 5.24, 5.27 **CSA CCM v4:** SEF-01, SEF-04 **PCI DSS 4.0:** 12.10.1, 12.10.2 **CIS Controls v8.1:** 17.1, 17.7

DOMAIN 8: GOVERNANCE AND CHANGE MANAGEMENT

Agents change constantly through model updates, prompt edits, and tool additions. Every one of those is a change to your control environment.

2 controls: AI-ASC-27, AI-ASC-28

AI-ASC-27 Model, Prompt, and Tool Version Control

Tier	Tier 2
Control statement	Model version, system prompt, tool definitions, and configuration are version controlled and change managed. Any change is tested before production and is attributable to a named approver.
Implementation guidance	Store prompts and tool definitions in source control with pull request review. Pin model versions explicitly and prohibit automatic upgrades in production. Run the adversarial test set before promoting any change. Record model version with every logged decision so behavior can be tied to configuration.
Risk if absent	A silent model or prompt change can alter agent behavior without any code deployment, defeating change control entirely. When behavior shifts, you cannot correlate it to a cause without version history.
Framework mapping	**NIST CSF 2.0:** PR.PS-01, ID.AM-08 **ISO 27001:2022:** 8.32, 8.9 **CSA CCM v4:** CCC-03, CCC-09 **PCI DSS 4.0:** 6.5.1, 6.5.2 **CIS Controls v8.1:** 4.2, 16.1

AI-ASC-28 Periodic Agent Risk Review

Tier	Tier 2
Control statement	Each production agent receives a documented risk review at least quarterly, covering permission drift, escalation trends, incident history, control effectiveness, and continued business justification.
Implementation guidance	Assign the review to the named agent owner with security sign-off. Compare current permissions against the last approved inventory and remediate drift. Review escalation and override statistics for evidence of rubber-stamping. Decommission agents without current business justification rather than leaving them idle.
Risk if absent	Agent permissions and scope expand incrementally, and each individual expansion looks reasonable. Without periodic review, an agent approved for a narrow task ends up with broad production access nobody consciously granted.
Framework mapping	**NIST CSF 2.0:** GV.OV-01, ID.RA-01 **ISO 27001:2022:** 5.35, 5.36 **CSA CCM v4:** GRC-06, IAM-02 **PCI DSS 4.0:** 12.4.2, 7.2.4 **CIS Controls v8.1:** 5.3, 18.1

PHASED DEPLOYMENT MODEL

Controls define what good looks like. Phasing defines how you get there without stopping the business. Do not skip phases, and do not shorten a phase because the agent is performing well. The failure modes that matter appear in weeks three and four.

Phase	Mode	What the agent does	Exit criteria
Phase 1	Observation only	Agent detects and recommends. Humans execute every action.	Minimum 30 days. Zero unexplained recommendations.
Phase 2	Low-risk automation	Agent handles enrichment, triage, and reversible actions.	Minimum 30 days. Escalation rate stable or declining.
Phase 3	Bounded decision automation	Agent executes within defined thresholds. Above threshold escalates.	Minimum 30 days. Zero unapproved threshold breaches.
Phase 4	Expanded autonomy	Thresholds widened based on demonstrated performance.	Entered only after 30 consecutive days of clean Phase 3 operation.

Regression rule

Any material change to model version, system prompt, tool set, or permission scope returns the agent to the prior phase for a minimum of 14 days. Change control that does not include a regression rule is documentation, not control.

FRAMEWORK CROSSWALK

Every control maps to at least one requirement in each of five major frameworks. Use this table to fold agent controls into an existing audit program rather than standing up a separate one.

ID	NIST CSF 2.0	ISO 27001:2022	CSA CCM v4	PCI DSS 4.0	CIS v8.1
AI-ASC-01	PR.IR-01, PR.AA-05	8.20, 8.22	IVS-03, IVS-07	1.2, 1.3, 1.4	12.2, 12.4, 4.4
AI-ASC-02	PR.AA-01, PR.DS-01	5.17, 8.24	IAM-06, CEK-03	3.5, 8.3.1	3.11, 5.2
AI-ASC-03	PR.AA-05, PR.PS-01	8.21, 8.26	AIS-02, IVS-06	1.3, 6.4.2	12.2, 16.10
AI-ASC-04	PR.IR-04, DE.CM-09	8.6	IVS-04, IPY-03	12.10.5	12.1, 13.1
AI-ASC-05	PR.PS-01, PR.PS-05	8.31, 8.19	IVS-09, CCC-04	2.2, 6.4.1	4.1, 4.6, 2.7
AI-ASC-06	PR.AA-05, PR.AA-01	5.15, 8.2	IAM-05, IAM-16	7.2, 7.3	6.8, 5.4
AI-ASC-07	ID.AM-01, GV.RR-02	5.9, 5.16	IAM-01, DCS-06	8.1, 12.5.1	1.1, 5.1
AI-ASC-08	ID.AM-02, GV.OC-04	5.9, 8.9	IAM-02, GRC-05	7.2.4, 12.5.2	1.1, 6.1
AI-ASC-09	PR.AA-01, PR.AA-02	5.17, 8.5	IAM-14, CEK-08	8.3.9, 8.6.3	5.2, 5.3
AI-ASC-10	PR.DS-02, PR.PS-06	8.26, 8.28	AIS-04, AIS-07	6.2.4, 6.4.1	16.10, 16.11
AI-ASC-11	PR.DS-01, ID.RA-09	5.23, 8.24	DSP-03, STA-06	6.4.3, 12.8	3.10, 15.4
AI-ASC-12	PR.PS-06, DE.CM-09	8.26, 8.28	AIS-04, TVM-02	6.2.4	16.11, 16.12
AI-ASC-13	DE.AE-02, DE.CM-01	8.16	LOG-13, TVM-03	10.4, 11.5	13.1, 8.11
AI-ASC-14	DE.AE-03, PR.PS-04	8.15	LOG-03, LOG-05	10.2.1, 10.2.2	8.2, 8.5
AI-ASC-15	PR.DS-01, DE.AE-06	8.15, 5.28	LOG-09, LOG-10	10.3.2, 10.5.1	8.3, 8.9
AI-ASC-16	RS.AN-03, DE.AE-02	5.28, 8.15	SEF-06, LOG-08	10.4.1, 12.10.1	8.11, 17.4
AI-ASC-17	ID.RA-05, DE.AE-04	8.16, 5.7	LOG-05, GRC-03	12.3.1	8.11
AI-ASC-18	RS.MA-02, PR.PS-04	8.14, 8.16	BCR-05, AIS-07	10.7.2, 12.10.5	17.3, 16.12
AI-ASC-19	GV.RR-01, PR.AA-05	5.3, 5.15	GRC-04, IAM-10	7.2.1, 12.5.2	6.8, 14.6
AI-ASC-20	RS.CO-02, GV.RR-02	5.24, 5.26	SEF-02, SEF-03	12.10.1, 12.10.3	17.2, 17.4
AI-ASC-21	RC.RP-01, RC.RP-03	8.13, 5.29	BCR-08, BCR-11	12.10.1	11.3, 11.5
AI-ASC-22	DE.CM-01, DE.AE-02	8.16	LOG-13, TVM-03	10.4.1, 10.6	8.11, 13.1
AI-ASC-23	DE.CM-09, PR.DS-06	8.26, 8.29	AIS-07, DSP-16	6.2.4, 11.6	16.11, 16.13
AI-ASC-24	DE.CM-01, PR.PS-04	8.16, 8.6	LOG-08, IVS-04	10.7.1, 10.7.2	8.11, 13.1
AI-ASC-25	RS.MI-01, RS.MI-02	5.26, 8.16	SEF-05, IAM-14	12.10.1, 8.2.6	17.5, 5.3
AI-ASC-26	RS.MA-01, ID.IM-02	5.24, 5.27	SEF-01, SEF-04	12.10.1, 12.10.2	17.1, 17.7

ID	NIST CSF 2.0	ISO 27001:2022	CSA CCM v4	PCI DSS 4.0	CIS v8.1
AI-ASC-27	PR.PS-01, ID.AM-08	8.32, 8.9	CCC-03, CCC-09	6.5.1, 6.5.2	4.2, 16.1
AI-ASC-28	GV.OV-01, ID.RA-01	5.35, 5.36	GRC-06, IAM-02	12.4.2, 7.2.4	5.3, 18.1

ABOUT OLYMPUS CYBER

Olympus Cyber is an executive cybersecurity firm specializing in incident response, digital forensics, and cyber resilience. We build structured security programs that translate technical risk into clear business decisions.

This library was written by people who run incident response for a living. The controls reflect what actually fails under pressure, not what reads well in a policy document.

Where this fits in an engagement

- Agent architecture review against all 28 controls, with a gap register and remediation sequence
- Adversarial testing of deployed agents including injection, privilege escalation, and failure handling
- Agent-specific incident response tabletop exercises for executive and technical teams
- Integration of agent controls into an existing SOC 2, ISO 27001, PCI DSS, or CMMC program

24/7 Response
response@olympus-cyber.com
+1-801-516-3347

Advisory
olympus-cyber.com
South Jordan, Utah

Executive Cyber Resilience

AI Agent Security Control Library v1.0 | August 2026 | Free to use with attribution.