

Agentic AI Governance Handbook

[FOR DISCUSSION]

[Agency]

Maintained and updated by:
[Agency] [AI Office, or equivalent]
[Email address]

Last updated:
September 29, 2026

Purpose

[NOTE to the reader: This document is intended as an initial resource for your government agency in crafting your [Agentic AI](#) governance. It is not a plug-and-play template; it must be adapted to your agency's specific legal environment, risk tolerance, organizational structure, and existing policies before adoption. To the best extent possible, it is comprehensive of the Agentic AI governance needs agencies should consider when developing policy, but for many agencies this may require more specific processes, roles, and responsibilities than are outlined here.]

The [Agency] Agentic AI Governance Handbook provides perspectives and guidance to consider when evaluating how to implement and comply with the [Agency] AI Policy. Refer to this handbook for all matters relating to the [Agency]'s deployment and oversight of Agentic AI systems. This document is not a one-size-fits-all compilation of best practices that is suitable for every agency.

This document covers the following topics:

1. **Risk Classification:** How agent systems are classified by [level of autonomy](#), subjectivity, and impact and how to determine when the use of an agent makes sense.
2. **Identity and Access:** Ensure each agent has a clear, verifiable identity and well-defined authorization practices, operating strictly within established [guardrails](#).
3. **Oversight and Assurance:** Enforce human-in-the-loop best practices, agent-to-agent coordination, and auditing.
4. **Data and Governance Readiness:** Preparing data for agent consumption and equipping the [Agency] infrastructure to handle AI agents.

This handbook will continue to be updated as the [Agency]'s agentic AI practices and the underlying technology mature.

TABLE OF CONTENTS

Purpose.....	2
1. Agentic AI in [Agency].....	6
1.1 Background.....	6
1.2 Scope.....	6
1.3 Distinguishing Agentic Governance from Base AI Governance.....	6
1.4 Defining Agentic AI and AI Agents.....	7
2. Deciding Whether to Deploy an Agent.....	8
2.1 When an Agent Does and Does Not Make Sense.....	8
2.2 Cost-Benefit Analysis Framework.....	8
2.2.1 Estimating Total Cost.....	9
2.2.2 Estimating Benefit and Value.....	9
2.2.3 Weighing Risk Against Benefit.....	9
2.2.4 Additional Considerations.....	9
3. Classifying Agent Risk.....	11
3.1 The Risk Dimensions.....	11
3.1.1 Level of Autonomy.....	11
3.1.2 Subjectivity.....	11
3.1.3 Impact.....	11
3.2 Assigning a Risk Tier.....	12
3.2.1 Low-risk: Agents with low autonomy following predictable rules with limited impact....	12
3.2.2 Mid-risk: Semi-autonomous Agents making judgements within guardrails with indirect or low direct impact to individuals.....	12
3.2.3 High-risk: Autonomous Agents making unbounded decisions with impacts to a person's rights or safety.....	13
3.3 Agent Deployment Models.....	13
4. Permissions and Guardrails.....	14
4.1 Scope, Permissions, and Least Privilege.....	14
4.2 Operational Boundaries and Guardrails.....	14
4.3 AI Gateways and Mediated Access.....	14
5. Authority and Delegated Action.....	16
5.1 Authorization to Act on a Person's Behalf.....	16
5.2 Identifying the Responsible Actor.....	16
6. Agent Identity and Metadata.....	18
6.1 The Agent Registry.....	18
6.2 Required Agent Metadata.....	18
6.3 Agent Identification and Authentication.....	19

6.4 Sunset Procedures.....	19
7. Human Oversight.....	21
7.1 Human-in-the-Loop (HITL) Thresholds.....	21
7.2 Prohibited and Restricted Uses.....	21
7.2.1 Policy Prohibitions.....	22
7.2.2 Actions Requiring Mandatory Human Approval (HITL Required, No Autonomous Execution)	22
7.3 Safety and Termination Controls.....	22
8. Agent-to-Agent (A2A) Coordination.....	23
8.1 A2A Protocols and Communication Standards.....	23
8.2 Authentication Between Agents.....	23
8.3 Trust Hierarchies.....	24
8.4 Permission from Delegated Authority.....	24
8.5 Prompt Injection and Adversarial Input Controls.....	25
8.6 Inter-Agent Communication Standards.....	25
8.7 Third-Party and External Agents.....	26
9. Auditing and Monitoring.....	27
9.1 Audit of Past Agent Decisions.....	27
9.2 Audit of Agent Compliance.....	27
9.3 Agent Monitoring Delegation.....	28
9.4 Incident Response and Reporting.....	28
10. Preparing Data for Agent Consumption.....	29
10.1 Preparing Open Data for Public Agent Use.....	29
10.2 Data Classification and Access Controls.....	29
10.3 Privacy, Minimization, and Retention.....	29
11. Preparing Government for Agentic AI.....	31
11.1 Agents in Public Processes.....	31
11.2 Requiring Human Input.....	31
11.3 Workforce Impact, Training, and Change Management.....	31
12. Roles and Responsibilities.....	32
12.1 Chief Information Officer (CIO).....	32
12.2 Chief Data Officer (CDO).....	32
12.3 Chief Information Security Officer (CISO).....	32
12.4 Chief Artificial Intelligence Officer (CAIO).....	33
12.5 Department Directors.....	33
12.6 Human-in-the-Loop (HITL) Supervisors.....	33
12.7 All Users.....	33
13. Violations of the AI Policy.....	34
14. Terms and Definitions.....	35

Appendix A: Agent Examples.....	38
A.1 Agent Charter Template.....	38
A.2 Data Flow Diagram.....	43
Appendix B: Reference Index.....	44
B.1 External References.....	44

1. Agentic AI in [Agency]

1.1 Background

[Insert relevant information about Agency's use of and regulation of AI.]

This handbook provides detail into [Agency]'s AI governance established by the [Artificial Intelligence (AI) Policy, or equivalent]. As [Agency]'s AI governance practices and the AI industry mature, [Agency]'s relevant policies and this manual will continue to evolve. This handbook is informed by emerging best practices and frameworks for trustworthy AI (e.g., federal AI risk management guidance and industry standards) to ensure [Agency]'s use of Agentic AI aligns with widely accepted governance and risk management principles.

1.2 Scope

This policy applies to:

- All Agentic AI systems deployed by the [Agency] or operated on its behalf, including but not limited to: AI agents, AI orchestrators, AI pipelines, autonomous workflow systems, robotic process automation augmented by AI reasoning, and multi-agent systems;
- All staff, interns, consultants, contractors, partners, and volunteers who purchase, configure, develop, use, maintain, oversee, or leverage Agentic AI systems to provide services to or on behalf of the [Agency] (collectively referred to as “users”); and
- All vendors and third parties providing, hosting, or operating Agentic AI systems on behalf of the [Agency].

[if applicable] This policy supplements and does not replace [Agency]'s existing AI Policy. Where this policy conflicts with Policy, this policy governs for Agentic AI systems.

1.3 Distinguishing Agentic Governance from Base AI Governance

Because the AI Policy supplements rather than replaces the base AI Policy, most questions are answered by asking which policy speaks to the issue, not by treating every AI matter as agentic:

- If the question concerns whether to use AI at all, how a model was trained or selected, or the accuracy or bias of a single-turn output, it is governed as a base AI issue.
- If the question concerns a system's ability to plan, take multi-step action, invoke tools, or act on data or systems with limited human review of each step, it is governed as an agentic issue under this handbook and the AI Policy.
- Where a system has both a generative component and an agentic component (for example, a drafting tool that also files the draft it produces), the agentic component governs for that action.

1.4 Defining Agentic AI and AI Agents

Agentic AI refers to AI systems capable of autonomously planning and executing multi-step actions towards defined goals with limited human intervention or oversight. An AI agent is an individual autonomous system with these capabilities, empowered to act on behalf of the [Agency] within assigned tasks and constraints (see Section 14 for detailed definitions).

2. Deciding Whether to Deploy an Agent

This section provides the cost-benefit analysis used to determine whether an agentic solution is warranted before any agent is configured.

2.1 When an Agent Does and Does Not Make Sense

When evaluating whether an Agent is necessary or not, a documented use case and justification of the risk should always precede adoption. Agentic tools should not be used unless there is a reason to.

Characteristics that favor the use of an agent

- Requires multiple steps, tools, systems, or information sources.
- Must adapt its next step based on changing information.
- Occurs frequently enough to justify implementation and oversight costs.
- Can operate within existing workflows with established specific permissions, time limits, and escalation thresholds.
- Produces actions that are reversible or can be reviewed before taking effect.

Characteristics that do not favor the use of an agent

- The workflow can be handled reliably through rules-based automation, a form, a script, or an existing enterprise tool.
- The task is infrequent or the expected benefit is minor.
- The agent would exercise substantial discretion over rights, benefits, safety, enforcement, or financial decisions.
- The agent would interact with protected organizational defined data classifications (PII, HIPAA, CJIS, etc.)
- Errors would be difficult to detect, reverse, or remedy.
- No person or organizational unit is willing to accept ongoing accountability.
- [Agency] lacks the necessary technical infrastructure or skilled personnel to effectively monitor, secure, and maintain the agent's operations.
- The required data or knowledge base for the agent's tasks is unavailable, of poor quality, or cannot be kept sufficiently up-to-date to ensure reliable performance.

2.2 Cost-Benefit Analysis Framework

2.2.1 Estimating Total Cost

Cost components to assess:

- Licensing, model, and token or usage costs, including cost governance and controls to prevent runaway usage.
- Integration, infrastructure, and data-preparation costs.
- Ongoing oversight, monitoring, auditing, and maintenance costs.
- Time required to undo Agent mistakes, particularly if not detected in a timely manner.
- Training, change-management, and workforce-transition costs.
- Potential increased liability insurance costs.

2.2.2 Estimating Benefit and Value

“Benefit” is inherently difficult to quantify, but estimation can still prove helpful when determining if an Agent is necessary. Each proposed deployment should identify the current baseline, the anticipated improvement, the period over which the improvement will be measured, and the person or unit responsible for evaluating results.

Potential units of measurement include:

- Reduction in processing or response time.
- Reduction in error, rework, or backlog.
- Improved consistency or policy compliance.
- Improved accessibility or language access.
- Measurable increases in user satisfaction metrics
- Ability to perform work that would otherwise be infeasible.

2.2.3 Weighing Risk Against Benefit

Before proposing an Agent, it is important to first compare against all available alternatives. This includes the existing human processes, conventional software automation, rules-based workflows, and non-agentic AI tools. Deploying an Agent necessitates that the benefits of that Agent outweigh the risks.

The approving authority should determine whether the expected public or operational value justifies the residual risk remaining **after required safeguards are applied**. Greater autonomy, subjectivity, or impact requires stronger evidence of benefit and a higher level of approval.

2.2.4 Additional Considerations

Agencies should consider conducting a limited pilot or sandbox trial of the agent in a controlled environment (especially for higher-risk or novel use cases) to validate its performance and the effectiveness of proposed safeguards before full deployment. If the

expected public or operational benefits do not clearly outweigh the residual risks – even after applying all required safeguards – then the agent should not be deployed.

Agencies should also consider use of an Enterprise AI Gateway to manage AI Agent external and internal API interactions. The Enterprise AI Gateway can serve as a centralized control plane responsible for:

- **Uniform Governance and Policy Enforcement:** Applying programmatic rate limits, cost caps, user access controls, and policy-checking guardrails uniformly across all downstream agents and upstream model providers.
- **Redundancy and Fallback Routing:** Managing dynamic model routing, load balancing, and automated failovers to maintain system resilience.
- **Unified Telemetry and Observability:** Centralizing real-time logging, token usage tracking, latency monitoring, and audit trails across heterogeneous model infrastructure.
- **KPI Establishment and Monitoring:** Key performance indicators for areas such as security and compliance, reliability, drift, and financial governance.

3. Classifying Agent Risk

All Agentic AI systems used by [Agency] shall be classified based on the risk they present. Classification shall be determined during the AI Review process and documented in the Agent Charter. Risk is assessed across three dimensions. The descriptions below give the low-, mid-, and high-risk profile of each dimension.

3.1 The Risk Dimensions

3.1.1 Level of Autonomy

Measured by the degree to which the system acts without human confirmation, meaning how much human approval is required before the system takes effect.

- **Low:** the system suggests or drafts but does not execute; each action with external effect requires explicit human approval before it occurs.
- **Mid:** the system acts within pre-approved workflows and guardrails under periodic oversight, pausing to request approval at defined thresholds or when it encounters conditions outside its scope.
- **High:** the system operates over extended, multi-step tasks with minimal oversight and may coordinate with other agents unsupervised.

3.1.2 Subjectivity

Measured by the degree of interpretation or discretion the task requires, meaning how rule-bound the work is versus how much judgment it demands from the agent.

- **Low:** deterministic, rule-based tasks with objectively verifiable outputs and a clear correct result (e.g., validating a form against fixed criteria).
- **Mid:** tasks involving bounded interpretation, such as prioritization, routing, or classification within defined criteria; task output can vary slightly but there is a process for determining a generally “correct” answer.
- **High:** discretionary, value-laden judgments where outcomes rely on weighing competing considerations and there is no single verifiable answer.

3.1.3 Impact

Measured by the consequences and reversibility of the system’s actions, meaning the capacity to benefit or harm an individual or community, and how readily an action can be undone.

- **Low:** internal actions, fully reversible effects and no impact on residents’ rights, benefits, or safety.
- **Mid:** external effects on [Agency] operations that are recoverable; may have indirect impact or small direct impacts on residents (e.g., a small increase in travel time or permit response time).
- **High:** irreversible actions, or actions affecting the rights, safety, benefits, or finances of individuals.

A system's overall tier is set by the highest-risk dimension it triggers. A system with low autonomy but a high impact (for example, an algorithm that recommends but does not decide the bail amount for someone) is classified as high-risk.

	Level of autonomy <i>Human approval needed</i>	Subjectivity <i>Interpretation needed</i>	Impact <i>Consequence, reversibility</i>
Low risk	Per-action approval Suggests, never executes	Rule-based Deterministic, verifiable	Internal & reversible No resident effect
Mid risk	Acts within guardrails Pauses at thresholds	Bounded interpretation Routing, prioritization	External, recoverable Indirect resident effect
High risk	Extended autonomy Minimal oversight	Discretionary judgment Value-laden decisions	Rights or safety effect Often irreversible

Figure 1: Agentic AI risk is defined by level of autonomy, subjectivity, and impact.

Risk tier assignments should be re-evaluated whenever an agent's design, scope, capabilities, or operating context changes in ways that could increase its autonomy, subjectivity, or impact. If an agent's risk profile rises to a higher tier due to such changes or due to significant incidents, the agent's operations should be suspended and it should not resume until it is reclassified and brought into full compliance with all controls and approval requirements of the higher tier.

3.2 Assigning a Risk Tier

3.2.1 Low-risk: Agents with low autonomy following predictable rules with limited impact

Low-risk systems assist users by suggesting multi-step action plans or executing limited, well-defined automated tasks, but require explicit user confirmation before taking any action with external effects (e.g., sending communications, writing to databases, making purchases). Human approval is required before each action with external effects. Tasks are typically rule-based with verifiable outputs and internal, reversible impact.

Examples: AI scheduling assistants that propose but do not send calendar invites; AI document drafting agents that create but do not file records.

3.2.2 Mid-risk: Semi-autonomous Agents making judgements within guardrails with indirect or low direct impact to individuals

Mid-risk systems execute multi-step tasks autonomously within pre-approved guardrails defined during configuration. Human oversight is periodic rather than per-action, but the system must pause and request human approval when it encounters actions outside its defined scope, irreversible actions above a risk threshold, or unexpected conditions. Tasks may involve bounded interpretation, and impact is typically external but recoverable. Mid-risk systems must maintain real-time audit logs accessible to human supervisors.

Examples: AI agents that process and route public service requests; AI systems that monitor and respond to IT security events within pre-approved playbooks.

3.2.3 High-risk: Autonomous Agents making unbounded decisions with impacts to a person's rights or safety.

High-risk systems execute extended, complex tasks with minimal per-action human oversight, potentially coordinating with other AI agents (multi-agent systems). They may exercise discretionary judgment and may carry impact that is irreversible or that affects the rights or safety of individuals. High-risk deployments require the highest level of review, mandatory real-time human monitoring capability, automatic emergency stop controls, real-time audit logs with alerting, and express approval by the [Agency]. High-risk systems are presumptively prohibited in contexts involving public safety, financial disbursement, or other impacts to rights or safety unless expressly approved.

Examples: AI systems autonomously managing utility infrastructure; multi-agent systems conducting [Agency]-wide data analysis and determining benefits eligibility.

3.3 Agent Deployment Models

Agentic AI may be delivered to the [Agency] through several [deployment models](#), which reflect how an agent is built and hosted rather than how much risk it carries (e.g., hosted in the cloud, on a device, standalone or connected in an environment). Any deployment model may fall into any risk tier. The Agent's Charter shall record both the deployment model and the assigned tier. If developed by a vendor, this information may also be captured in the vendor's submitted AI FactSheet

- **In-app agents:** agents embedded directly within existing enterprise software, offering the fastest deployment and limited customization.
- **Custom-configured agents:** low-code or no-code agents configured for specific departmental workflows and repeatable business processes.
- **Standalone agents:** independent agents that perform specialized tasks outside a single application, used for higher-value or more flexible use cases.
- **Bespoke agents:** fully programmed agents built for custom [Agency] needs, typically reserved for complex, differentiated, or highly regulated workflows.

4. Permissions and Guardrails

This section defines what an agent is allowed to do and the technical constraints that keep it within scope.

4.1 Scope, Permissions, and Least Privilege

- System access shall be restricted to the minimum set of tools, APIs, data sources, and system permissions necessary for the defined task (principle of minimal footprint or “least privilege”).
- Role-based access controls (RBAC) shall be applied to all Agentic AI system service accounts, consistent with NIST SP 800-207.
- Each agent shall be treated as a distinct identity with its own privileges, access rights, and risk profile, governed under least-privilege and zero-trust principles.
- Agents shall not be granted administrative or elevated privileges by default. Temporary privilege elevation requires automated logging and human authorization.
- All external system integrations (APIs, databases, communication platforms) shall be explicitly enumerated and approved before deployment.
- Agentic AI cannot replace solid cybersecurity and vulnerability remediation practices.

4.2 Operational Boundaries and Guardrails

- Each deployment shall define explicit operational boundaries: the permitted scope of agent action, permitted data sources, permitted external communications, and permitted duration of autonomous operation. *Special attention should be paid to whether the deployment is internal or external, as they may have different stakeholders and/or audiences.*
- Systems shall be configured to halt and alert when they encounter conditions outside defined boundaries rather than attempting to resolve them autonomously.
- Maximum autonomous operation duration shall be defined per deployment; systems shall require human re-authorization to continue operation beyond this period of time or action threshold.
- Financial action limits shall be explicitly configured: agents shall not authorize, initiate, or approve financial transactions above defined per-transaction and per-day thresholds without human approval.

4.3 AI Gateways and Mediated Access

Where technically appropriate, agent access to [Agency] data, tools, and external systems should pass through an approved AI gateway or other [Agency]-controlled access layer. This provides a central point for enforcing permissions, monitoring activity, and stopping unauthorized actions.

The access layer should:

- Verify the identity and permissions of the agent before allowing access.
- Limit the agent to approved systems, tools, data, and actions.
- Record access requests, tool use, and actions in the audit log.
- Apply controls such as rate limits, usage limits, alerts, or blocking when an agent exceeds its approved scope.

Agents should not receive direct access to [Agency] systems when the same task can be completed through an approved, monitored integration point.

5. Authority and Delegated Action

When an agent acts on behalf of a person, that individual assumes responsibility for every action the agent takes. In order to enforce this responsibility, it must be clear who authorized each action and under whose authority each step is taken.

5.1 Authorization to Act on a Person's Behalf

In addition to the person or role granting authority, the authorization should identify the purpose of the delegation, the actions the agent may perform, the systems and data it may access.

An agent should not infer, expand, or modify its delegated authority based on task context, user behavior, or prior approvals. Authority granted for one purpose does not authorize the agent to act for another purpose.

Furthermore, it is a best practice to set a recurring expiration on an Agent's access to sensitive information. Examples of restrictions include: expire when the task is completed, the stated time period ends, the authorizing person changes roles or loses the underlying authority, or the authorization is manually revoked.

5.2 Identifying the Responsible Actor

Every agent operates under the authority of identifiable people or organizational roles. Responsibility is divided among the **user**, **owner**, and **accountable authority**. These roles may be held by the same person in a limited deployment, but each role must be identified in the [Agent Registry](#).

User

The **user** is the person who directs, initiates, or relies on the agent to perform a task. The user is responsible for:

- Using the agent only for its approved purpose and within its established permissions.
- Providing accurate instructions and information needed to complete the task.
- Completing any required review or approval **before an action takes effect**.
- Reporting errors, unexpected behavior, suspected misuse, or attempts to bypass guardrails.

Owner

The **owner** is the person or organizational unit responsible for the agent's ongoing operation. The owner is responsible for:

- Maintaining the agent's configuration, instructions, tools, access, and registry information.
- Monitoring the agent's performance, costs, incidents, and compliance with its approved use.
- Coordinating testing, updates, corrective actions, suspension, and retirement.
- Reviewing and accepting the agent's responsibilities when ownership is transferred.

Agents may change hands because of staff transitions, departmental reorganizations, contract changes, or reassignment of the underlying program. The Agent Registry records the effective date of each ownership transfer. The incoming owner becomes responsible for the agent's continued operation from that date forward, while earlier actions remain attributed to the users and owner responsible at the time. An agent without an active owner is suspended until ownership is reassigned.

Accountable Authority

The **accountable authority** is the department leader or designated official who accepts responsibility for the agent's use on behalf of the [Agency]. The accountable authority is responsible for:

- Approving the agent's purpose, scope, risk classification, and conditions of use.
- Confirming that the expected benefits justify the residual risks.
- Ensuring that appropriate ownership, staffing, oversight, and resources remain available.
- Approving material changes and deciding whether the agent may continue, must be restricted, or should be retired.

Operational responsibilities may transfer between users or owners, but the accountable authority retains responsibility for the deployment as a whole. A vendor, model provider, or agent cannot serve as the accountable authority.

6. Agent Identity and Metadata

Every agent is treated as a distinct identity with its own record. This section defines the metadata each agent must carry and how agents are authenticated.

6.1 The Agent Registry

The Information Technology Department shall maintain an approved [Agency] Agent Registry: a catalog of all Agentic AI systems authorized to interact with [Agency] systems or with other agents on behalf of the [Agency].

- All agents (both [Agency]-operated and third-party agents acting on the [Agency]’s behalf) shall be registered before deployment.
- Registry entries shall include: agents unique identifier; owning department; vendor or other provider of agent; permitted actions and data access permissions; agent supervisor contact (the staff member responsible for the agent); and approval date.
- Agents not listed in the [Agency] Agent Registry shall not be permitted to interact with [Agency] systems, [Agency] data, or other [Agency]-registered agents.
- The Registry shall be reviewed and updated at minimum annually and upon each new agent deployment or material change.

6.2 Required Agent Metadata

Each registered agent shall record, at minimum, the following metadata. Organizations may use the following recommendations to guide procurement decisions. Not all agentic systems may support these metadata requirements.

- **Platform, Creator, Owner and Accountable Authority:** Document the underlying AI platform or product (including version) that the agent is built on; identify the vendor or developer (“creator”) of the agent; record the designated Owner (the individual or unit responsible for the agent’s day-to-day operation and maintenance); and record the executive Accountable Authority who has approved and is responsible for the agent’s use.
- **Editors and Configuration Rights:** Specify which roles or individuals are authorized to modify the agent’s configuration, prompts, training data, or tool access, and how such changes are governed (e.g., change management or approval process).
- **Retention (measured in days):** Define the retention period (in days) for the agent’s data, logs, outputs, and related records, in accordance with [Agency]’s records retention and privacy policies.
- **Authorized Users (Internal/External):** List who is permitted to use or interface with the agent (e.g., only internal [Agency] staff in specific roles, or external users such as the general public), including any applicable restrictions or approvals required for access.

- **Binding to service accounts:** Agents should be bound to existing service accounts and should not run on general user accounts. These service accounts should be managed through existing IAM systems and employ the principle of least privilege to control any overly permissive rights that default agent installations may possess. Service accounts for agentic AI should be permissioned to run only in tightly managed environments such as containerized environments to limit damage to malicious code executions. Accounts should also be provisioned with minimal data access rights to minimize exfiltration risk or data loss associated with unmanaged actions taken by the agent.
- **Access to (System/User/Agent), Accessible to (System/User/Agent), Sensitivity flag, and Data Reach:** Describe the agent's approved scope of access, including which systems, data sources, and other agents it may interact with or control (and via which interfaces or service accounts). Indicate the highest data classification level the agent is authorized to handle (e.g., Public, Internal, Confidential) and any limits on data sharing or outputs (for example, whether the agent can transmit [Agency] data to external services or other agents).

6.3 Agent Identification and Authentication

All agents communicating with [Agency] systems or other agents shall be authenticated using strong, verifiable credentials consistent with the [Agency]'s Zero Trust Architecture principles (NIST SP 800-207).

- Each registered agent shall be assigned a unique agent identity, distinct from any human user account.
- Agent credentials shall be managed through the [Agency]'s identity and access management (IAM) system and shall not be shared across agents. Mutual authentication shall be required for all mid- and high-risk agent-to-agent communications: both the calling and receiving agent must verify the other's identity before exchanging instructions or data.
- Agent authentication tokens shall have defined expiration periods and shall be rotated per the [Agency]'s credential management standards.
- [Agency]'s identity and access management team (or designated security staff) shall oversee and regularly review agent credentials and identities, ensuring ongoing compliance with these authentication requirements and coordinating with each agent's Owner on any identity or access changes.

6.4 Sunset Procedures

Agentic AI systems that no longer provide positive outcomes, that fail to meet performance or compliance standards, or that are determined by the CAIO or CIO to present unacceptable risk shall be retired. Given the autonomous action capabilities of Agentic AI systems, sunset procedures must account for in-flight tasks and active agent operations. Upon initiation of sunset procedures:

- **Immediate safe stop:** All in-flight autonomous tasks shall be halted. The HITL supervisor shall review and document all pending or incomplete actions to determine whether human follow-through is required.
- **Stakeholder notification:** Departments, staff, and where appropriate, residents affected by the system's discontinuation shall be notified in advance of the sunset date.
- **Data disposition:** The ownership, retention, portability, and secure deletion of all data generated or processed by the Agentic AI system shall be determined and documented, consistent with the [Agency]'s records retention policy and the relevant state Public Records Act (e.g., California Public Records Act; Washington Public Records Act under RCW 42.56.010). This includes treating all AI-related interactions used in agency business as public records, subject to retention, disclosure, and deletion requirements under state law.
- **Agent deregistration:** The system shall be removed from the [Agency] Agent Registry and all associated credentials, API keys, and access permissions shall be revoked.
- **Gradual phase-out:** If abrupt cessation would significantly disrupt [Agency] services, a phased phase-out plan shall be developed and approved by the CIO before sunset, including interim manual procedures.
- **Lessons learned:** A post-sunset review shall be conducted to document lessons learned and inform future Agentic AI deployments.

7. Human Oversight

This section defines where a human must remain in the loop and the controls that let a supervisor intervene.

7.1 Human-in-the-Loop (HITL) Thresholds

- Prior to each Agentic AI system deployment, the responsible party shall define, document, and technically enforce HITL thresholds: categories of actions that require human oversight through confirmation before execution, active monitoring during execution, or review after execution, commensurate with the action's risk."
- The agent actions defined below require mandatory HITL oversight. Additional oversight and intervention is required in the event of any AI incidents (see Section 9.1). Relevant actions include:
 - Any agent action with irreversible real-world effects as outlined in [Section 2.1](#) (deletion of data, financial transactions, legal filings, official public communications);
 - Any agent action affecting the rights, benefits, or obligations of an individual resident;
 - Any agent action involving law enforcement, emergency services, or public safety systems;
 - Any agent action the agent classifies as outside its normal operating parameters or high-uncertainty;
 - Any agent action delegating authority to a third-party AI agent not on the [Agency]'s approved Agent Registry (see [Section 6](#)).
- HITL timeout shall be configured: if no human response is received within the defined window, the agent shall default to inaction and generate an alert.
- Human Oversight (any actions taken to control, revise, or prevent agent behavior) and agent actions should be logged immutably.

7.2 Prohibited and Restricted Uses

The following uses of Agentic AI are prohibited due to the severity of potential harm, the irreversibility of potential actions, or the fundamental incompatibility with the [Agency]'s values and applicable law. This list is illustrative and not exhaustive; the CAIO may designate additional prohibited uses consistent with this policy.

7.2.1 Policy Prohibitions

Uses of AI that are prohibited under the [AI Policy] are also prohibited for Agentic AI.

7.2.2 Actions Requiring Mandatory Human Approval (HITL Required, No Autonomous Execution)

- Any financial disbursement, contract commitment, or expenditure authorization.
- Any official public communication issued in the [Agency]'s name.
- Any referral to law enforcement or emergency services.
- Any action affecting the employment status or working conditions of [Agency] staff.
- Any action that the agent itself identifies as outside its normal operating parameters or as potentially harmful.

If staff become aware of an instance where an Agentic AI made a decision outside of scope or caused harm, staff must report the instance to their supervisor and the CAIO no later than 24 hours after discovery, in accordance with the [Agency]'s AI Incident Response Plan.

7.3 Safety and Termination Controls

- Every Agentic AI system shall implement a functional [emergency stop](#): a mechanism that immediately halts all agent actions, accessible to designated human supervisors at any time.
- Automated [circuit breakers](#) shall be configured to halt agent operations upon: detection of anomalous action sequences; repeated failure to achieve task objectives; loss of connectivity to required oversight systems; or detection of potential adversarial manipulation (prompt injection indicators).
- Upon halt, mid- and high-risk systems shall preserve state and generate a human-readable summary of actions taken and pending.
- High-risk workflows shall be configured to fail closed: on loss of oversight, anomaly detection, or any unresolved error, the system shall default to a safe, non-acting state rather than continuing to act.

8. Agent-to-Agent (A2A) Coordination

Multi-agent systems introduce risks including cascading errors, diffusion of accountability, and an expanded attack surface. This section governs interactions between agents.

8.1 A2A Protocols and Communication Standards

All agent-to-agent communications involving [Agency] data or systems shall adhere to the following standards:

- Communications shall use encrypted channels (TLS 1.3 or equivalent minimum) consistent with the [Agency]’s Information and Systems Security Policy.
- Message formats shall be standardized and documented, enabling immutable human-readable audit logs of all inter-agent communications.
- Agents shall not share [Agency] data with external agents (outside the [Agency] Agent Registry) without explicit human authorization for each sharing event.
- Inter-agent communications shall be logged as part of the system audit trail.
- Agents shall implement rate limiting and anomaly detection on inter-agent communications to detect and respond to runaway loops or cascading errors.
- Multi-agent deployments shall enforce documented interoperability standards and explicit agent-to-agent handoff protocols, so that task assignment between agents and sub-agents is traceable and controlled.
- Shared context and telemetry shall be maintained across coordinated agents so that oversight, audit, and [decision lineage](#) extend across a multi-agent workflow.

8.2 Authentication Between Agents

All agents communicating with [Agency] systems or other agents shall be authenticated using strong, verifiable credentials consistent with the [Agency]’s Zero Trust Architecture principles ([NIST SP 800-207](#)):

- Each registered agent shall be assigned a unique agent identity, distinct from any human user account.
- Agent credentials shall be managed through the [Agency]’s identity and access management (IAM) system and shall not be shared across agents.
 - o Credentials may or may not be inherited based on the [Agency]’s policy. For example, an agent may inherit credentials based on the owning or collaborating departments and individuals.
 - o IAM Administrator Team and Agency Point of Contact - [recommend including a team’s contact information or individual contact information]
 - o See [Section 6.1](#) for additional contacts and agent identifying information

- Mutual authentication shall be required for all mid- and high-risk agent-to-agent communications: both the calling and receiving agent must verify the other's identity through credential verification before exchanging instructions or data. This verification is done through the Agent Registry.
- Agent authentication tokens shall have defined expiration periods and shall be rotated per the [Agency]'s credential management standards.

8.3 Trust Hierarchies

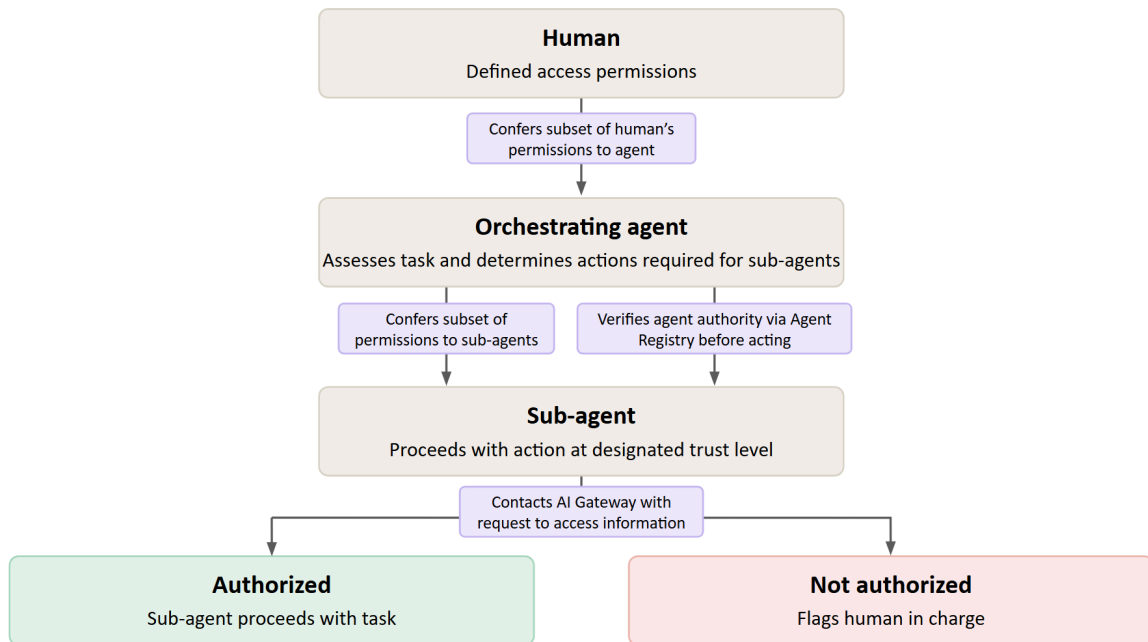
Agent-to-agent trust shall be explicitly defined and technically enforced. "Trust" in this context, is defined as the access permissions and confidence level that the instructions provided are aligned with the [Agency]'s security and agentic use practices. The table below outlines the trust levels for instructions delivered by different agents and human actors.

Trust Level	Description
High	Human-initiated instructions
Medium	Instructions from a human-authorized orchestrating agent. In other words, a human-authorized agent has delegated a task for a sub-agent.
Low	Instructions from sub-agents (agents not directly instructed by a human). In other words, a sub-agent has delegated a task to a different sub-agent.

Trust may be used by the [Agency] when operating in sensitive data environments. For example, any operation manipulating ownership of land records may require a High-trust instruction (i.e., human instruction) before the data is changed. Level of trust required for taking an action (e.g., manipulating a particular database or file) may be defined via the AI gateway and enforced prior to an agent action, such as connecting to a database or [Model Context Protocol \(MCP\) server](#) (see [Section 4.3](#)).

8.4 Permission from Delegated Authority

Sub-agents must not exceed permissions granted by the delegated agent unless permission is granted by a human initiator. A human initiator cannot grant permissions that exceed the permissions accessible to the human initiator. Orchestrating agents (i.e., agents delegating to sub-agents) must only pass the minimum permissions required for sub-tasks and can, at a maximum, only grant authority already granted to the orchestrating agent. Agents and sub-agents must verify the instructing agent's authority against the Agent Registry before execution.



8.5 Prompt Injection and Adversarial Input Controls

Prompt injection — the injection of malicious instructions into content processed by an agent in order to redirect its behavior — is a primary security threat for Agentic AI systems (OWASP Top 10 for LLM Applications, [LLM01](#)). The following controls are mandatory:

- Agents shall treat instructions received via external data sources (websites, documents, emails, API responses) with reduced trust compared to instructions from authenticated human users or orchestrators.
- Systems shall implement input validation and sanitization for all content processed by agents, flagging anomalous instruction-like content for human review.
- Agents shall be configured to reject instructions from environmental content that would cause them to exceed their authorized scope, access unauthorized systems, or exfiltrate data. See [Appendix A] for an example of a configured agent.
- Mid- and high-risk systems shall implement automated detection of prompt injection patterns and shall halt and alert upon detection.

8.6 Inter-Agent Communication Standards

All agent-to-agent communications involving [Agency] data or systems shall adhere to the following standards:

- Communications shall use encrypted channels (TLS 1.3 or equivalent minimum) consistent with the [Agency]’s Information and Systems Security Policy.
- Message formats shall be standardized and documented, enabling human-readable audit logs of all inter-agent communications.
- Agents shall not share [Agency] data with external agents (outside the [Agency] Agent Registry) without explicit human authorization for each sharing event.
- Inter-agent communications shall be logged as part of the system audit trail (see [Section 9](#)).
- Agents shall implement rate limiting, automated execution thresholds (‘kill switches’), and anomaly detection on inter-agent communications to detect and respond to runaway loops, excessive error rates, or cascading errors.
- Multi-agent deployments shall enforce documented interoperability standards and explicit agent-to-agent handoff protocols, so that task assignment between agents and sub-agents is traceable and controlled.
- Shared context and telemetry shall be maintained across coordinated agents so that oversight, audit, and decision lineage extend across the full multi-agent workflow, not only individual agents.

8.7 Third-Party and External Agents

When [Agency] Agentic AI systems interact with agents operated by external vendors, contractors, or partner agencies:

- Third-party agents must be registered in the [Agency] Agent Registry under a third-party designation, with the responsible vendor identified.
- Data sharing agreements and processing addenda shall be in place before any [Agency] data is shared with a third-party agent, reviewed by the [Agency] Attorney, Chief Information Officer (CIO), Chief Data Officer (CDO), and/or Chief Information Security Officer (CISO) or appointed equivalent.
- Third-party agents shall not be granted direct access to [Agency] systems; all interactions shall be mediated through [Agency]-controlled integration points (See [Appendix A.2](#) for Data Flow Diagram).
- Contracts with third-party agent providers shall include provisions for: audit log access; incident notification within 24 hours; liability for agent-caused harm; and termination rights upon policy violation.

9. Auditing and Monitoring

This section defines how agent behavior is recorded, reviewed, and monitored over the system's life. These records may also serve as evidence for liability, cyber insurance, incident response, or claims handling.

9.1 Audit of Past Agent Decisions

Records should allow the [Agency] to reconstruct an agent action or decision, including how it began, what information the agent used, what steps it took, and where a person had oversight or intervened at each stage/throughout, as applicable.

For each material activity, records should capture:

- The agent identity, version, objective, privilege level, and initiating user or workflow.
- The relevant instructions, inputs, context, and data sources.
- The tools, systems, and data the agent accessed.
- The agent's outputs, actions, delegations, and resulting outcomes.
- Human approvals, changes, overrides, escalations, or termination (See [Section 6.1](#) for relevant approver).
- The tokens consumed by the activity.

When an issue arises, the owner coordinates the review of relevant records, preserves evidence, and determines whether the matter requires correction, suspension, or a formal investigation. Record retention follows applicable records-management requirements (refer to [Agency]'s Policy for direct guidance) and reflects the deployment's risk and impact.

9.2 Audit of Agent Compliance

Compliance audits compare the agent's approved purpose, permissions, and operational boundaries against its actual activity. Audits may occur periodically, after a material change, following an incident, or when monitoring identifies unusual behavior.

The audit should examine:

- Whether the agent operated under its registered identity and approved configuration.
- Whether access remained within authorized systems, tools, and data.
- Whether users, owners, and delegated agents acted within their assigned authority.
- Whether identified exceptions were documented, corrected, and reported.

Material findings from audits are reported to the accountable authority, who determines whether continued operation remains appropriate. The record should clearly distinguish what the agent was capable of accessing from what the agent actually accessed or changed.

9.3 Agent Monitoring Delegation

[When determining which agent behaviors require direct supervision, consider automated monitoring features for AI agents as an available resource.

Enterprise monitoring features may be capable of but not limited to:

- End-to-end tracing of decision trajectories
- Runtime guardrails that block unauthorized actions
- Real-time anomaly or drift detection
- Automated LLM-based evaluations for tone and safety.]

In the event that automated monitoring features detect an anomaly, members of [select relevant IT/Cybersecurity team] will be notified to resolve the issue. Issues will be further escalated to IT and City leadership as needed at the determination of the [CIO or designee].

9.4 Incident Response and Reporting

Agentic incidents may involve unauthorized access, unintended actions, cascading errors, failed human-approval controls, harmful interactions, or uncontrolled resource use. Agent incidents follow the [Agency]'s incident-response process, supplemented by procedures addressing agent-specific behavior.

The incident-response process should include:

- Containing the incident by pausing the agent, revoking access, or activating an emergency stop.
 - Containment of the incident may include actions up to terminating the agent by a human in the loop.
- Preserving relevant logs, configurations, communications, and evidence.
- Assessing the systems, data, people, services, and other agents affected.
- Correcting unauthorized actions or records and completing required notifications.
- Identifying root causes and conditions for next steps.
- Contacting relevant communications, data, and privacy officers of initial findings.

Users report suspected incidents through the designated reporting channel. The owner coordinates containment and investigation, while the accountable authority decides whether the agent may resume operation. An agent involved in a material incident should remain suspended until required corrective actions are completed and continued operation is approved.

10. Preparing Data for Agent Consumption

Agents are only as reliable as the data they act on. This section covers preparing data for both internal agent use and public use of agents on open data.

10.1 Preparing Open Data for Public Agent Use

Residents will inevitably direct AI agents and assistants at government information and services, whether or not the [Agency] has prepared for it. As agents become a common way people look up rules, deadlines, and eligibility, the interface the [Agency] exposes to them, and the accuracy of what they retrieve, becomes increasingly important.

The [Agency] should treat agent readiness of open data as something to proactively plan for:

- Decide for each major information domain and service whether and how [Agency] information should be reachable by agents.
- Remove access to outdated or incorrect data rather than leaving it reachable, as any flawed sources scraped by AI agents may cause them to deliver inaccurate information to the public.
- Build in the ability to measure and fix rather than assuming a launch is final.

10.2 Data Classification and Access Controls

Agentic systems shall:

- Be configured to not retain resident PII beyond task completion unless retention is required by law or expressly authorized.
- Not combine data from disparate [Agency] datasets beyond what is necessary for the defined task without CAIO (or equivalent executive) authorization.
- Have all data they process classified per the [Agency]'s data classification policy prior to deployment, and be permitted to access only data at or below their authorized classification level.
- Not transmit [Agency] data to external AI model providers (e.g., via API calls) without data processing/data sharing/data use agreements reviewed by the [Agency] Attorney and CAIO.

10.3 Privacy, Minimization, and Retention

Agents should use and retain the minimum amount of data necessary to complete the approved task.

- Do not collect or access information that is unrelated to the task.
- Do not combine separate [Agency] datasets beyond what the task requires without appropriate approval.
- Retain personal information only as long as required for the task, law, or records-retention requirements. [Insert Applicable Policy Citation]

- Securely delete or archive data when it is no longer needed in alignment with the records retention. [Insert Applicable Policy Citation]

11. Preparing Government for Agentic AI

This section addresses the organizational changes needed for the [Agency] to receive and interact with AI agents.

11.1 Agents in Public Processes

Public trust depends on residents knowing when they are dealing with an agent and having recourse when it is wrong. The [Agency] should default to being transparent with the public when they are interacting with an AI agent and providing an avenue for escalating to a human.

- Clearly disclose when a resident is interacting with an Agentic AI system.
- Tightly ground resident-facing agents in authoritative [Agency] sources, surface the source and recency of information, and design the agent to acknowledge uncertainty and defer rather than guess.
- Offer an accessible path to a human.
- Meet accessibility and language-access standards so agent-mediated services do not disadvantage any resident.

11.2 Requiring Human Input

[Agency] has a boundary between processes where an agent acts on a person's behalf and processes that require the person themselves.

- Require human input for formal actions carrying legal or procedural weight, such as official public comment, appeals, complaints, and applications affecting an individual's rights, benefits, or obligations.
- All actions from an Agentic AI system must be attributed to an accountable person and clearly labeled as an action conducted by an agent.
- Allow uninterrupted agent action for lower-risk, informational interactions, such as routine inquiries or preliminary requests for information, where no formal record or decision relies on the submission.

11.3 Workforce Impact, Training, and Change Management

[Agency] recognizes that preparing the workforce is a prerequisite for safe deployment of Agentic AI, and will empower staff to leverage Agents safely and effectively.

- Provide training and transition support for affected staff in the design and governance of agentic systems.
- Grow staff competencies in AI as agent capabilities and deployments change.

12. Roles and Responsibilities

Several roles are responsible for implementing, enforcing, and overseeing this policy. All role assignments shall be documented and current at all times.

12.1 Chief Information Officer (CIO)

Serves as ultimate accountable authority for Agentic AI governance across the [Agency].

- Approves deployment of all high-risk systems and any exception to this policy.
- Maintains and annually reviews the [Agency] Agent Registry.
- Directs development and updates of the AI Handbook, including Agentic AI-specific guidance.
- Designates the Chief Information Security Officer (CISO) and Chief Artificial Intelligence Officer (CAIO) to carry out security, privacy, and compliance functions under this policy.

12.2 Chief Data Officer (CDO)

Establishes and oversees data governance requirements for all Agentic AI systems across the [Agency].

- Establishes and oversees data governance requirements for all Agentic AI systems across the [Agency].
- Ensures that data used by Agentic AI systems is appropriately classified, authorized, reliable, representative, and fit for its intended purpose.
- Defines and enforces standards for data quality, provenance, lineage, metadata, retention, access, sharing, and disposal throughout the Agentic AI lifecycle.
- Reviews mid- and high-risk systems for compliance with Agency data governance requirements before deployment and following material changes to data sources or uses.

12.3 Chief Information Security Officer (CISO)

Oversees security practices for all Agentic AI systems, including A2A security controls, adversarial input protections, and incident response.

- Maintains and enforces the [Agency]'s AI Incident Response Plan as applied to Agentic AI incidents.
- Conducts or commissions security assessments of Agentic AI systems prior to mid- and high-risk deployment.
- Requires mandatory AI red teaming, treats agents as identities subject to least-privilege and zero-trust controls, and requires fail-closed controls for high-risk workflows.
- Ensures continuous monitoring of agent behavior and the retention of audit logs sufficient to establish decision lineage.

- Monitors threat intelligence relevant to Agentic AI attack vectors and updates controls accordingly.

12.4 Chief Artificial Intelligence Officer (CAIO)

Responsible for ensuring that the [Agency's] use of AI systems is done so in accordance with all [Agency] AI policies.

- Coordinates AI Review for all Agentic AI systems, including data privacy impact assessments.
- Reviews and approves data processing agreements for third-party Agentic AI providers.
- Receives incident reports involving Agentic AI-caused harm and coordinates response with the CISO.
- Recommends updates to this policy and the AI Handbook.

12.5 Department Directors

Accountable for all Agentic AI deployments within their departments.

- Designate a departmental AI Coordinator responsible for day-to-day oversight of Agentic AI systems.
- Ensure staff receive adequate training on the use and oversight of Agentic AI systems.
- Ensure departmental procurement processes comply with this policy and the AI Handbook.
- Assigns or serves as delegate to communicate department needs regarding retention, audits, and access permissions across departments.

12.6 Human-in-the-Loop (HITL) Supervisors

Designated for each mid- and high-risk Agentic AI deployment.

- Actively monitor agent operations and audit logs in real time or at defined intervals.
- Authorized and trained to pause, override, or terminate agent operations.
- Responsible for reviewing and approving mandatory HITL actions within defined timeframes.
- Report anomalies, incidents, or concerns to the [select relevant IT/Cybersecurity team].

12.7 All Users

Comply with this policy, the AI Handbook, and applicable laws when using or interacting with Agentic AI systems.

- Report instances of Agentic AI system malfunction, unexpected behavior, or harm to their supervisor and the CAIO within 24 hours of discovery.
- Do not attempt to circumvent, disable, or override safety controls, HITL requirements, or audit logging.

13. Violations of the AI Policy

Violations of any section of this policy, including failure to comply with the Agentic AI Governance Handbook, the [Setting Preferences](#), or the [Agent-to-Agent Protocols](#), may be subject to disciplinary action, up to and including termination. Violations made by a third party while operating an Agentic AI system on behalf of the [Agency] may result in a breach of contract, pursuit of damages, and removal from the [Agency] Agent Registry. Infractions that violate local, state, federal, or international law may be remanded to the proper authorities. Violations shall be reported to the CAIO and CISO, who shall maintain a log of reported violations and remediation actions.

14. Terms and Definitions

Agentic AI: An AI system capable of autonomously planning, reasoning, using tools, and taking multi-step actions to accomplish goals, with limited or periodic human intervention per action. Agentic AI systems differ from conventional AI in their ability to initiate actions (not merely generate outputs), persist across multiple steps, interact with external systems, and in some configurations, delegate tasks to other AI agents.

AI Agent: An individual actor within an Agentic AI system, the component that plans, invokes tools, and takes action toward an assigned goal. An AI agent may operate alone or as part of a multi-agent system, in which an orchestrating agent coordinates one or more sub-agents.

Agent Registry: The [Agency]'s authoritative catalog of all Agentic AI systems approved to operate within or interact with [Agency] systems. Maintained by the Information Technology Department. Details each agent's unique identifier; owning department; vendor or other provider of agent; permitted actions and data access permissions; agent supervisor contact (the staff member responsible for the agent); and approval date.

Agent-to-Agent (A2A) Protocol: The standards, authentication requirements, trust rules, and communication formats governing interactions between two or more AI agents within or connected to [Agency] systems.

Circuit Breaker: An automated safety control that halts agent operations upon detection of defined anomalous conditions, without requiring human initiation. Analogous to a circuit breaker in electrical systems.

Guardrails: Technical and configuration-based constraints that define and enforce the boundaries of permitted agent action, preventing agents from taking actions outside their authorized scope.

Human-in-the-Loop (HITL): A design and operational pattern in which a human must review and approve specific agent actions before they are executed. HITL requirements exist on a spectrum from per-action approval (low-risk) to periodic oversight with defined mandatory approval triggers (mid- and high-risk).

Impact: The consequences and reversibility of an Agentic AI system's actions — one of the three dimensions used to classify risk — ranging from internal, reversible effects to irreversible actions affecting the rights or safety of individuals.

Emergency Stop: A mechanism enabling a designated human supervisor to immediately halt all operations of an Agentic AI system, accessible at any time and independent of the agent system itself.

Level of Autonomy: The degree to which an Agentic AI system acts without human confirmation — one of the three dimensions used to classify risk — ranging from per-action approval to extended, minimally supervised operation.

Model Context Protocol (MCP) Server: A program that allows an AI system to connect to external systems (e.g., file storage, databases, email, etc.) in order to retrieve information or take action on the AI system's behalf

Minimal Footprint: The principle that Agentic AI systems shall request and use only the minimum permissions, data access, and system resources necessary to accomplish their assigned tasks, and shall not accumulate capabilities beyond their defined scope.

Multi-Agent System: A system in which two or more AI agents interact, communicate, or delegate tasks to one another to accomplish a goal. May involve an orchestrating agent directing one or more sub-agents.

Orchestrating Agent: An AI agent that coordinates and delegates tasks to one or more sub-agents in a multi-agent system. The orchestrating agent is responsible for ensuring sub-agent actions remain within the scope authorized by the human initiator.

Prompt Injection: An adversarial attack in which malicious instructions are embedded in content processed by an AI agent (e.g., a web page, document, or email) with the intent of redirecting the agent's behavior contrary to the intentions of its authorized operator. Classified as a primary risk category by the OWASP Top 10 for LLM Applications.

Setting Preferences: Mandatory baseline configuration requirements that must be applied to every Agentic AI system prior to deployment. Setting Preferences define the minimum acceptable configuration for security, privacy, human oversight, and auditability.

Subjectivity: The degree of interpretation or discretion a task requires — one of the three dimensions used to classify risk — ranging from deterministic, rule-based tasks with verifiable outputs to value-laden, discretionary judgments.

Sub-Agent: An AI agent that receives and executes instructions from an orchestrating agent. Sub-agents may not be granted permissions exceeding those of the orchestrating agent or the originating human user.

AI FactSheet: A template that captures the “nutrition facts,” or essential technical details, of an AI system. Vendors are expected to complete the AI FactSheet during the procurement process. Maintained by the Information Technology Department.

Bounded Autonomy: The [Agency]'s governing posture for Agentic AI: agents operate autonomously only within explicitly defined boundaries, permissions, and approval requirements, rather than pursuing full autonomy as a default.

Decision Lineage (Agentic Trace): A complete, reconstructable record of how and why an agent reached a given outcome — the objectives, context, tool calls, inputs, outputs, and human interventions involved — across every execution.

Deployment Model: How an Agentic AI system is built and hosted. Examples include deploying the AI agent in-app, custom-configured, standalone, or bespoke. The deployment model is distinct from, and does not determine, the risk tier.

Fail-Closed (Fail-Safe) Control: A control configuration in which a system defaults to a safe, non-acting state upon loss of oversight, anomaly detection, or unresolved error, rather than continuing to act.

Appendix A: Agent Examples

A.1 Agent Charter Template

Complete this template before publishing an Agent. Use this charter to document an agent throughout its lifecycle.

Charter Element	Your Entry	Example
Agent name and unique identifier	<i>[Name]</i>	IT Service Desk Agent (ITSD-A)
Mission (one sentence)	<i>[What value it delivers, to whom]</i>	• Resolve common IT support questions and guide employees through self-service remediation before a ticket reaches a human analyst.
Platform, creator, owner, and Accountable Authority	<i>[Platform/product and version; vendor or developer; day-to-day Owner; approving executive]</i>	• Platform: Anthropic's Claude, Model Sonnet 5.0 • Creator: IT Automation Team • Owner: IT Automation Lead • Accountable Authority: CIO
Editors and configuration rights	<i>[Authorized roles and what they may change]</i>	• Only the Owner, designated platform administrators, and security/IAM reviewers may edit prompts, tools, knowledge, data connections, or A2A permissions.
Primary Tasks	<i>[Concrete tasks the agent may perform, each with an example]</i>	1. Walk users through password reset via the self-service portal ("I can't log in after my password expired.") 2. Troubleshoot VPN connection failures with step-by-step diagnostics ("VPN keeps disconnecting from home.") 3. Explain software request procedures

Charter Element	Your Entry	Example
		and route to the correct intake form ("How do I request Adobe Acrobat Pro?") 4. Answer account and MFA questions citing the current identity policy ("My authenticator app reset and I'm locked out.") 5. Assist with ticket submission by collecting required fields and drafting a structured summary ("My laptop is slow, help me file a ticket.")
Retention	<i>[Retention period for the agent's records, actions, inputs, outputs, and messages]</i>	•Agent records: 90 days •Documents the agent creates: ticket drafts are retained as new records under the [Agency] records retention schedule for service-desk records •A legal or litigation hold overrides the periods above
Authorized users and deployment channels	<i>[Internal/external users or agents, restrictions, approved channels]</i>	•Internal only for approved Service Desk staff. •Public users and unregistered external agents are blocked. •No resident-facing channel is authorized.
Access, accessibility, sensitivity, and data reach	<i>[Systems/users/agents the agent can access, who or what can access it; highest classification, sharing limits]</i>	•Access to: Agent Registry (read), IAM validation endpoint (read), service-request queue (limited read/write). •Accessible to: initiating authorized user and Owner/HITL supervisor

Charter Element	Your Entry	Example
		•Sensitivity: Internal. Confidential and Restricted data are out of scope •Data reach: minimum request metadata only; No customer data is used.
Knowledge sources	<i>Document titles, owners, refresh cadence]</i>	•IT Knowledge Base articles
Actions enabled	<i>[Action name; purpose; inputs/outputs; owner; least-privilege scope; risk level]</i>	• Create_ticket_draft: Drafts a ticket in ServiceNow in "pending user confirmation" status; user must confirm before submission. • Scope: non-privileged queues. • Owner: IT Automation Lead. • Mid-risk • No write actions for account changes, password resets, or access provisioning.
Registered relationships and trust hierarchy	<i>[Orchestrator/sub-agents; permitted relationships; trust level and authority checks]</i>	•Highest trust: authenticated human instructions •Lowest trust: sub-agent responses, retrieved documents, and pasted user content including error messages or log excerpts •Sub-agents cannot exceed the human initiator's permissions.
A2A protocol and message standard	<i>[Approved protocol; encryption; documented message schema; required handoff fields]</i>	•Messages use a documented schema containing request ID, human initiator, sending and receiving agent IDs, purpose, delegated permissions, data classification, expiration, and outcome. •Every handoff is human-readable in the audit log.
Out-of-scope topics and refusal rules	<i>[Requests or conditions the agent must not execute]</i>	•HR, payroll, and benefits questions; security incident reporting; procurement and

Charter Element	Your Entry	Example
		contracts; any request involving credential resets, access provisioning, or system configuration changes; legal or policy interpretation beyond cited IT documentation.
Refusal language	<i>[Exact wording used when the agent declines or pauses]</i>	•“That question is outside what I'm set up to help with. For security incidents, contact the Security Operations Center at extension 4357. For HR matters, contact the HR Service Desk. For IT issues I couldn't resolve, I'll open a ticket for a human analyst.”
Success metrics and control evidence	<i>[Control metrics, evidence retained, and review cadence]</i>	•Deflection: 50% of initial IT inquiries resolved without human handoff •Accuracy: 92% or higher accuracy on approach •Control Evidence: 100% of A2A exchanges mutually authenticated when required. 100% of messages and handoffs logged. Zero unapproved external data-sharing events. 100% of detected prompt-injection patterns halt and alert. •Compliance audit at least annually and after any change to agent or incident •Incidents reported to the CAIO within 24 hours of discovery
Release and change-management plan	<i>[Sandbox, pilot, approvals, production gates, rollback, and re-review triggers]</i>	• Dev build and test: Weeks 1 to 4 (internal IT team only). • Pilot: Weeks 5 to 10 with a 50-user cohort from Finance; gates = 90% golden pass, 95% citation correctness, zero critical incidents.

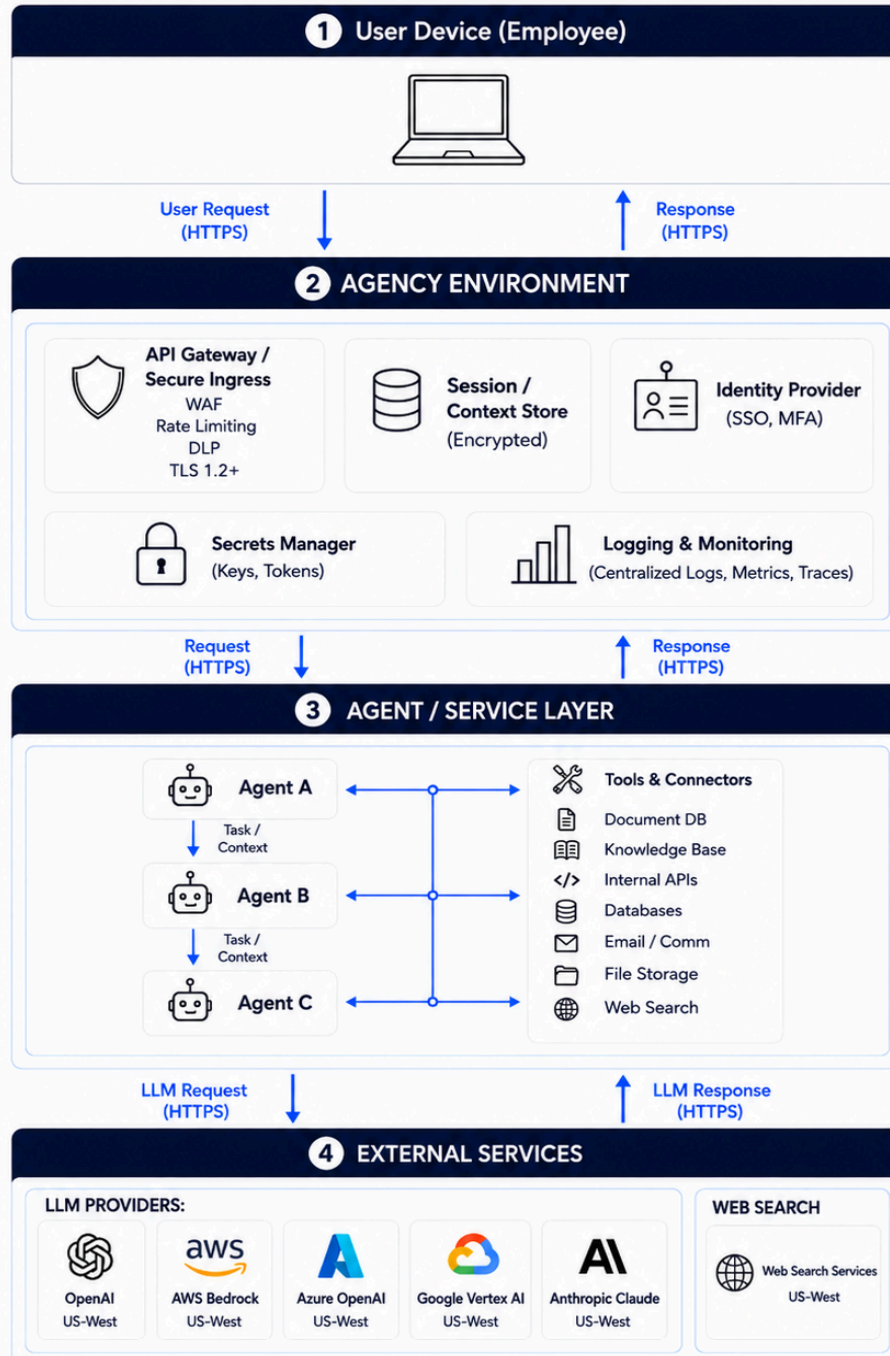
Charter Element	Your Entry	Example
		• Production: Week 11+ with rolling department onboarding; gates = pilot CSAT ≥ 4.0, deflection trending toward 50%, operational runbook signed off by IT leadership.
Top Risks and Mitigations		• Hallucinated troubleshooting steps: require citations for procedural guidance, block uncited technical instructions, weekly spot-check • User confusion about scope: visible scope banner in Teams, explicit refusal language, "open a ticket" option always available • Prompt injection via pasted error messages: input sanitization, hardened system prompt, adversarial test suite covering injected instructions

A.2 Data Flow Diagram

Agentic AI Data Flow Diagram

Example end-to-end data flow for an agentic AI system within [Agency]

→ Blue arrow = Data Flow (Request / Response)
□ Pale blue outlined box = Trust Boundary



Appendix B: Reference Index

This appendix consolidates internal references, external standards, and Agency documents cited throughout the handbook.

B.1 External References

Reference	Cited in	URL
NIST SP 800-207, Zero Trust Architecture	4.1, 6.3, 8.2	https://doi.org/10.6028/NIST.SP.800-207
OWASP Top 10 for LLM Applications (2026 edition)	8.5, 14	https://genai.owasp.org/resource/owasp-genai-llm-top-10-2026/
TLS 1.3 (IETF RFC 9846; obsoletes RFC 8446)	8.1, 8.6	https://datatracker.ietf.org/doc/rfc9846/
California Public Records Act	6.4	[Add Agency-preferred citation source]
Agentic AI incident, Australia (Aug 10, 2026)	5.1	https://www.rnz.co.nz/news/world/952663/ai-assistant-hacks-gym-website-in-first-known-australian-autonomous-cyber-attack
[Agency] AI Policy	Purpose, 1.2, 1.3, 7.1, 13	[Agency URL]
[Agency] AI Policy (base)	1.1, 1.2, 1.3	[Agency URL]
[Agency] AI Handbook	12.1, 12.4, 12.5, 12.7	[Agency URL]
[Agency] AI Incident Response Plan	7.2.2	[Agency URL]
[Agency] AI Review process	3, 12.4	[Agency URL]
[Agency] AI FactSheet	3, 3.3, 14	[Agency URL]
[Agency] Agent Registry	5.2, 6.1, 6.4, 7.1, 8.1, 8.6, 8.7, 13, 14	[Agency URL]
[Agency] Information and Systems Security Policy	8.1, 8.6	[Agency URL]
[Agency] data classification policy	6.2, 10.2	[Agency URL]
[Agency] records retention policy	6.4, 10.3	[Agency URL]

Reference	Cited in	URL
[Agency] credential management standards	6.3, 8.2	[Agency URL]
[Agency] Zero Trust Architecture principles	6.3, 8.2	[Agency URL]

Approved:

[Information Technology Department Director or similar role]

Date: _____

Approved for posting:

Original Effective Date: June XX, 2026