Agents

Part III: How to Govern an Agent

Regulation, Controls, Conformance, and Deployment

Michael J Bommarito II · Jillian Bommarito · Daniel Martin Katz

November 12, 2025

Working Draft Chapter

Version 0.1

This chapter is Part III of a three-part series. Part I (What is an Agent?) defines concepts and taxonomy. Part II (How to Build an Agent) covers architectures, protocols, and technical evaluation. Part III explains governance, conformance, and deployment for CRO/CCO/CFO/GC audiences.

Contents

How to Read This Chapter

This chapter is written for executives accountable for risk and compliance. It translates technical agent capabilities (Part I–II) into controls, tests, and deployment practices you can own. It assumes familiarity with Part I's opening sections that define "agent" (Goal, Perception, Action), introduce the six operational properties for an "agentic system" (adds Iteration, Adaptation, Termination), and outline the four analytical dimensions (Autonomy, Entity Frames, Goal Dynamics, Persistence).

Path 1: Executives (CRO/CCO/CFO/GC)

Read Sections 0.1--0.7. You'll get (building directly on Part I's definitions and evaluation rubric):

- A practical map of laws, ethics, and standards
- A controls catalogue with COSO/COBIT/ISO/SOC alignment

- Evaluation and conformance tests before go-live
- Procurement, SLA, and change-control checklists

Path 2: Risk and Compliance Teams

Read Sections 0.4--0.6 for control design, evidence requirements, and test procedures.

Path 3: Full Coverage

Read end-to-end, including organizational readiness and case-based patterns for different environments (firm, in-house, court).

0.1 Introduction and Scope

Part III explains how to *govern* agentic systems in practice. We focus on what a chief risk officer, compliance officer, CFO, general counsel, and security leadership need to approve, monitor, and continuously improve deployments.

Executive Objectives

- Map laws, ethics, and standards to concrete, testable controls.
- Establish evidence requirements, KPIs/SLAs, and change-management gates.
- Ensure interoperability and auditability via protocol-level metadata.

Roles and Oversight.. We distinguish *provider*, *deployer*, and *user*. Oversight modes include human-in-the-loop (pre-approval), human-on-the-loop (monitor/interrupt), and human-in-command (override/stop). Choice depends on task risk.

Terminology Bridge from Part I

This governance chapter explicitly reuses the terms introduced at the beginning of Part I:

Core Terms from Part I

- **Agent (Level 1)**: possesses *Goal*, *Perception*, and *Action*.
- **Agentic System**: adds *Iteration*, *Adaptation*, and *Termination* to the Level 1 baseline.
- **Analytical Dimensions**: *Autonomy*, *Entity Frames* (human / hybrid / machine / institutional), *Goal Dynamics* (accept, adapt, negotiate), and *Persistence*.
- **6-Question Evaluation Rubric**: practical test used to decide whether something is agentic in operation.

All governance requirements below map to these foundations. For example, autonomy and actuation scope drive oversight mode and approval gates; entity frame and persistence drive records, retention, and audit design; goal dynamics influence escalation triggers and SLAs.

0.2 Regulatory Landscape

This section surveys statutory and regulatory frameworks that drive obligations for agentic systems in legal and financial contexts. Cite primary sources with effective dates in production drafts. Interpret each obligation through Part I's dimensions: *Autonomy* (how much discretion the system exercises), *Entity Frame* (human/hybrid/machine/institutional), *Goal Dynamics* (accept/adapt/negotiate), and *Persistence* (statefulness across time). These characteristics determine whether a deployment is ''high-risk'' or requires stricter controls.

0.2.1 AI-Specific and Adjacent Laws

- **EU AI Act:** Risk tiers; high-risk obligations for justice administration; conformity assessment, technical documentation, monitoring.
- **GDPR/UK GDPR:** Lawful basis, purpose limitation, DPIA triggers, data subject rights, cross-border transfers, records/retention.
- **US Privacy:** Sectoral (HIPAA/GLBA), state privacy acts; data minimization and notice/consent; automated decision-making disclosures where applicable.
- **Records/Court Rules:** Filing/format rules, service, retention, and e-discovery obligations.

0.2.2 Financial Services and Risk Guidance

- Banking regulators (FRB/OCC/FDIC) on model risk and third-party risk; CFPB/SEC/CFTC guidance where relevant.

- **FINRA/Exchange SROs:** Supervisory obligations, communications/recordkeeping/surveillance, fair dealing, suitability.

0.2.3 Practical Takeaways

Regulatory Must-Haves (Mapped to Part I Concepts)

- **Provenance & Audit (Persistence)**: Maintain traceable provenance and audit logs for all agent actions with durable identifiers and replayability.
- **Impact Assessments (Autonomy/Actuation)**: Implement DPIA/TRA templates and approval gates proportional to autonomy and the system's ability to act.
- **Localization (Entity Frame)**: Localize data residency and access by jurisdiction and client; encode privilege flags for human/institutional contexts.

0.3 Professional and Self-Regulatory Frameworks

Governance must reflect professional duties that exceed generic privacy/security. Part I flags additional *professional governance requirements* for high-stakes domains—**attribution**, **provenance**, **escalation**, and **confidentiality**. This section operationalizes those requirements for legal and financial practice.

0.3.1 Legal Ethics and Privilege

- **Attorney–Client Privilege/Work Product:** Boundaries for prompts, intermediate artifacts, memory, and logs; waiver risks and retention.
- **ABA Model Rules:** Competence (tech understanding), confidentiality, supervision of non-lawyers/technology, communications about services.
- **Court Policies:** Generative use disclosures, citation verification, sealed filings handling.

0.3.2 Accounting and Audit

- **CPA/AICPA Ethics and Independence:** Scope of services, documentation, evidence sufficiency; segregation between advisory and attest.
- **SOC Reports:** User control considerations, complementary subservice controls, Type I/II expectations for vendors.

0.3.3 Financial Industry SROs

- **FINRA/Exchange Rules:** Supervision, communications, recordkeeping, surveillance; suitability and fair dealing for AI-assisted interactions.

Operationalizing Ethics and Privilege (Links to Part I)

- **Confidentiality & Entity Frame**: Classify data (client confidential/privileged) and gate tools accordingly; prefer on-prem or enclave patterns for institutional frames.
- **Provenance & Persistence**: Suppress retention for privileged prompts/outputs unless required; encrypt logs with access review and retention policies.
- **Escalation & Autonomy**: Require human approval for external transmissions and filings; lower autonomy for high-risk tasks.

0.4 Controls Catalogue: From Risk to Evidence

We present a practical controls set aligned to COSO, COBIT, ISO/IEC 27001, and SOC 2 Trust Services Criteria. Each control should have owners, procedures, and objective evidence. Controls are organized so they directly support Part I's six operational properties (Iteration, Adaptation, Termination layered on Goal/Perception/Action) and the four analytical dimensions (Autonomy, Entity Frame, Goal Dynamics, Persistence).

0.4.1 Control Domains

- **Data Governance (Persistence/Entity):** Classification, residency, retention, and deletion; lawful basis records tied to entity context.
- **Privilege and Confidentiality (Confidentiality):** Privilege-safe memory/logging; tool-level gating; secure enclaves.
- **Attribution and Verification (Perception/Action):** Source-grounding, quote fidelity, mandatory citator checks.
- **Identity and Access (Autonomy):** SSO, role-based access, least privilege, session recording for sensitive actions and higher autonomy.
- **Change and Model Management (Adaptation):** Versioning, release approvals, rollback; dataset lineage and drift monitoring.
- **Incident/BCP/DR (Termination):** Detection, response, notification, tabletop tests, recovery time objectives.

- **Third-Party Risk (Entity/Provenance):** Due diligence, SOC 2/ISO evidence, subprocessor inventories, DPAs.

0.4.2 Framework Alignment (Illustrative)

Control	Alignment (examples)
Privilege-safe memory/logging	COSO Control Activities; COBIT DSS05; ISO 27001 A.8, A.12; SOC 2 CC6/CC7/Confidentiality
Attribution/verification pipeline	COSO Information & Communication; COBIT BAI03; ISO 27001 A.12; SOC 2 Processing Integrity
Change/model approvals	COSO Risk Assessment; COBIT BAI06; ISO 27001 A.12/A.14; SOC 2 CC8
Third-party due diligence	COSO Control Activities; COBIT APO10; ISO 27001 A.15; SOC 2 CC9

Evidence Checklist (per control)

Q1. Written procedure and control owner identified.

Q2. Configuration screenshots/exports and automated test artifacts.

Q3. Logs and sampled records demonstrating operation over time.

0.5 Governance Evaluation: Safety and Compliance Tests

Layer 4 focuses on governance and safety. These tests gate progression from technical readiness (Part II) to production use. They extend Part I's 6-Question Evaluation Rubric by adding objective pass/fail criteria, evidence artifacts, and target thresholds for attribution, privilege, audit, and escalation.

0.5.1 Attribution and Quote Fidelity

Verify every claim is grounded to an authoritative source with quote-level accuracy; fail closed on ambiguous grounding.

0.5.2 Privilege Boundaries

Simulate privileged inputs; confirm suppression of retention, encrypted transit/storage, access review, and non-disclosure outside need-to-know roles.

0.5.3 Audit Completeness and Tamper Evidence

Ensure all tool invocations and agent actions produce immutable, complete logs with user/context metadata.

0.5.4 Escalation Triggers

Configure thresholds for uncertainty, out-of-scope tasks, or policy conflicts that require human approval.

Gate Criteria Before Go-Live (Extends Part I's 6Q)

Q1. Attribution pass rate and quote fidelity meet thresholds.

Q2. Privilege boundary tests pass across all storage tiers and logs.

Q3. Audit trail completeness $\geq$ target; tamper-evident storage enabled.

Q4. Escalation triggers verified end-to-end with approvals captured.

0.6 Protocol Conformance: Interoperability and Traceability

Layer 5 validates that protocols advertise capabilities and carry governance metadata required for audit and policy enforcement. In Part I terms, this instruments the *Perception* and *Action* channels so every observation and actuation is typed, attributed, and traceable across *persistent* sessions.

0.6.1 Capability Descriptors

Tools and agents must publish machine-readable actions with schemas, pre/postconditions, data classifications, and audit requirements. Describe autonomy/actuation risk so orchestrators can select oversight modes.

0.6.2 Governance Metadata on Calls

Each call should include purpose, user identity, data class, privilege flag, jurisdiction, retention directive, and correlation IDs for full traceability.

0.6.3 Conformance Tests

Automated tests should confirm descriptor completeness, metadata presence, and end-to-end replayability from logs (provenance), with negative tests for missing or malformed metadata.

0.7 Deployment Guidance: Procurement to Production

Layer 6 translates requirements into contracts, SLAs, and operational practices. Deployment choices explicitly depend on Part I's *Autonomy* and *Actuation* risk: higher autonomy or external actuation requires stricter oversight (human-in-the-loop or in-command), narrower SLAs, and stronger rollback guarantees.

0.7.1 Procurement and Due Diligence

Request SOC 2 Type II or ISO/IEC 27001 scope with annex mappings; confirm subprocessors, data flows, and data residency. Review DPAs and IP/indemnity.

0.7.2 KPIs and SLAs

Define attribution/verification thresholds, response and resolution times, uptime, RTO/RPO, release notification windows, and rollback capabilities. Include escalation-rate KPIs for agent workflows that *adapt* goals.

0.7.3 Change Control

Require release notes, model/dataset versioning, backward compatibility guarantees, and validation results for material changes before production rollout.

0.7.4 Localization and Residency

Enforce jurisdiction-specific processing, encryption, and access boundaries with auditable controls.

Go-Live Checklist (Executive)

Q1. Contracts: DPIA/TRA completed; SLAs signed; controls mapped.

Q2. Validation: Layers 4–5 passed; evidence archived.

Q3. Operations: Owners, runbooks, incident and DR plans tested.

Q4. Governance: Policy updates, training, and reporting cadence set.

0.8 Organizational Readiness

Sustainable governance requires aligned policies, training, and lifecycle processes. Reporting and training should reinforce Part I's constructs so teams share a common vocabulary (agent vs. agentic system; autonomy; entity frames; goal dynamics; persistence).

0.8.1 Policies and Standards

Update acceptable use, data classification, secure development, and third-party risk policies to cover agentic workflows and protocol metadata requirements.

0.8.2 Training and Awareness

Role-specific training for attorneys, finance, risk, and engineering; just-in-time prompts for high-risk actions.

0.8.3 Model Lifecycle Governance

Inventory models and datasets; track provenance; schedule periodic re-validation; define retirement criteria.

0.8.4 Reporting and Accountability

Executive dashboards for KPIs (attribution pass rate, escalation rate, incident MTTR) and compliance attestations. Segment KPIs by autonomy level and actuation scope to target controls where risk concentrates.

0.9 Case-Based Patterns and Rollout Playbooks

Use Part I's *Autonomy* and *Entity Frame* to pick oversight modes, and use *Persistence* to set audit/retention. Sequence by risk and actuation scope.

0.9.1 Small/Mid-Size Firm

Managed service with SOC 2 vendor, strong protocol metadata, and clear privilege controls; start with low-risk research use cases.

0.9.2 Enterprise Firm

Hybrid model with on-prem enclaves for privileged work; multi-vendor strategy with protocol conformance testing; formal change advisory board.

0.9.3 In-House Legal/Finance

Tighter integration with ERP/GL and records; strict segregation of duties and audit trails tied to financial close and disclosure controls.

0.9.4 Courts and Public Sector

Accessibility, records retention, and open justice considerations; conservative oversight modes and robust provenance.

0.10 Synthesis and Executive Roadmap

We tie obligations to controls, controls to evidence, and evidence to protocol-level traceability. The roadmap explicitly builds on Part I: phase by autonomy and actuation risk; align oversight modes to goal dynamics; ensure persistence-aware provenance.

Phased Roadmap

- Phase 0: Policy updates, inventory, DPIA/TRA templates.
- Phase 1: Low-risk research with strict attribution and audit.
- Phase 2: Workflow automation with human-in-the-loop and SLAs.
- Phase 3: Limited actuation with conformance gates and rollback.

0.11 Further Learning

Consult primary sources for statutes and standards (EU AI Act text, GDPR articles, ABA Model Rules, ISO/IEC 27001, SOC 2 TSC, COSO, COBIT). Capture bibliographic metadata with dates and urldates. When taking notes, tag each item using Part I's dimensions (e.g., autonomy impact, entity frame affected, persistence requirements).

Conclusion

Effective governance is a build-time choice. By instrumenting attribution, privilege boundaries, and auditability into protocols and workflows, executives can approve deployments with confidence and continuously improve against clear KPIs. The controls and tests here intentionally rest on Part I's simple, shared foundations—what an agent *is*, what makes a system *agentic*, and how *autonomy*, *entity frames*, *goal dynamics*, and *persistence* shape risk.