Governance Maturity Model

v2.0.0

Author: John J. McCormick · Updated: February 28, 2026 · CC BY 4.0 · aiagentgovernance.org

SUGGESTED CITATION

McCormick, J. J. "Governance Maturity Model for AI Agent Operations." v2.0.0, February 2026. aiagentgovernance.org. https://aiagentgovernance.org/framework/maturity-model/

> ***Authorship context:*** *This is a practitioner's methodology, not an academic paper. The author built and operated the governed agent system described here — writing governance controls in production as agent failures occurred, documenting incidents in real time, and extracting the methodology from operational experience. The maturity model was derived from the drift detection progression observed in those operations, with regulatory mapping applied subsequently.*

Abstract

The Governance Maturity Model defines five levels of organizational readiness for AI agent governance. Organizations at Level 0 deploy agents without formal oversight structures. Organizations at Level 4 set the governance standard that others measure against. The model provides a self-assessment framework for determining current maturity, identifying gaps, and charting a path toward robust agent governance. Each level builds on the capabilities of the previous level — there are no shortcuts.

---

1. Why a Maturity Model

Most organizations deploying AI agents today have no formal governance methodology. They may have model-level compliance (bias testing, fairness metrics), security controls (guardrails, tool-use restrictions), and observability (tracing, logging). None of these address the operational governance question: *Are our agents doing the right work, the right way, with appropriate human oversight?*

The maturity model provides a structured answer to "where do we start?" and "what does good look like?" It maps the progression from ungoverned agent deployment to governance-as-a-discipline, with specific capabilities at each level.

The model was derived from two sources: the drift detection maturity progression observed in production multi-agent operations, and the governance capability requirements identified through regulatory mapping against the EU AI Act, NIST AI RMF, and ISO 42001.

Governance Authority Progression

Orthogonal to the five maturity levels, governance systems progress through three authority structures:

Authority Level	Description	Maturity Alignment
Advisory	Policies exist; compliance is expected but not structurally enforced	Level 0–1
Enforced	Execution requires structural authorization; agents cannot advance work without passing defined gates	Level 1–2
Separated Authority	Multiple independent actors required for authorization; no single actor can both propose and approve	Level 2–4

The authority progression is independent of — but complementary to — the maturity levels. An organization may be at Level 1 maturity with Enforced authority, or at Level 2 maturity with Advisory authority in some dimensions. Both the maturity level and the authority level should be assessed.

---

2. The Five Levels

Level 0: Ungoverned

State: Agents deployed without formal governance. No phase gates. No behavioral monitoring. No structured audit trail.

Characteristics:

- Agents execute tasks based on direct instructions with no lifecycle management
- Work output reviewed only when problems become visible
- No structured incident documentation
- Agent autonomy is implicit — not explicitly scoped or calibrated
- No distinction between agent types, roles, or trust levels

Risk profile: Maximum exposure. Behavioral drift is undetectable. Failures are discovered after impact, if at all. The organization cannot demonstrate governance to regulators, auditors, or customers.

Regulatory posture: Organizations at Level 0 are unlikely to meet the operational governance requirements of the EU AI Act (for systems subject to high-risk classification under Annex III), NIST AI RMF, or ISO 42001. Whether a specific AI agent deployment falls within these frameworks depends on deployment context and must be assessed for each implementation. Independent legal review is required before any compliance determination. See §5 for regulatory mapping guidance.

Where most organizations are today.

Level 1: Reactive Governance

State: Basic governance structures in place. Drift is detected after impact through human observation or post-incident review.

Capabilities introduced:

- **Phase-gated lifecycle.** Agent work follows defined phases (Plan, Build, Review, Approve) with explicit transitions
- **Human-in-the-loop.** A human operator approves phase transitions before agents proceed
- **Basic behavioral awareness.** The organization recognizes a set of known failure patterns (see the Behavioral Pattern Taxonomy (patterns.md)) and monitors for them manually
- **Incident documentation.** Deviations from expected behavior are logged, even when outcomes are acceptable
- **Audit trail.** Phase transitions, approvals, and gate decisions are recorded

Risk profile: Failures are detected after impact but documented and analyzed. The incident log becomes the primary input for governance improvement. Behavioral drift is visible in retrospect.

Regulatory posture: Addresses core operational requirements for human oversight (EU AI Act Article 14), basic record-keeping (Article 12), and initial risk management (Article 9). Provides baseline evidence for NIST AI RMF GOVERN and MAP functions. Organizations should conduct independent legal review before relying on these mappings for compliance purposes.

Independence constraint: At Level 1, a single human operator sits at every phase gate. This means that one cognitive failure mode (automation bias, attention drift, fatigue) can compromise multiple defense layers simultaneously. Principle 2 (Defense Layers Must Fail Independently) provides partial independence at this level, not full independence. This is a known and managed limitation — it is the structural rationale for progressing to Level 2, where independent validation functions introduce genuine defense layer separation.

Key gap: Detection is reactive. The operator is the sole detector, and detection occurs after impact.

Level 2: Structural Governance

State: Automated mechanisms detect drift at the point of action, before impact. Governance checks operate in the work pipeline.

Capabilities introduced:

- **Automated behavioral monitoring.** Known behavioral patterns are detected automatically during agent execution, not just during post-incident review
- **Prerequisite validation.** Phase transitions are checked against plan requirements — agents cannot skip steps or bypass prerequisites
- **Independent validation.** A dedicated validation function (separate from the executing agent) verifies compliance before human approval
- **Structured impact statements.** Every phase transition includes a structured risk assessment (risk level, blast radius, reversibility) for the approving human
- **Trust calibration.** Agent autonomy is informed by behavioral history — agents with higher drift rates receive closer oversight

Risk profile: Known failure modes are detected at the point of action. The validation function provides a compliance-oriented perspective that complements the operator's quality-oriented perspective (see MOIAS methodology (governance-lifecycle.md), Section 3.4 on OODA orientation diversity).

Regulatory posture: Addresses core operational governance requirements of EU AI Act Articles 9, 12, and 14 for high-risk systems. Note: Article 13 (provider-to-deployer transparency obligations) is a supply-chain obligation owed by AI system providers to deployers and is not addressed by this operational governance methodology. Maps primarily to NIST AI RMF GOVERN function (GV-2: accountability structures, GV-5: policies and controls, GV-6: oversight mechanisms); MAP and MANAGE coverage is

partial. Provides evidence toward ISO 42001 certification. Organizations should conduct independent legal review — particularly sub-article analysis of Article 9 — before relying on these mappings for compliance purposes.

Key gap: Detection is limited to known patterns. Novel failure modes are not caught until they become incidents.

Level 3: Predictive Governance

State: Trend analysis across validation logs, incident patterns, and behavioral metrics identifies emerging risk before a specific incident occurs. The system detects acceleration toward a safety boundary, not just crossing a boundary.

Capabilities introduced:

- **Trend detection.** Time-series analysis of validation results, incident frequency, and behavioral metrics identifies drift trajectories before they produce incidents
- **Predictive risk assessment.** Trust calibration identifies agents trending toward failure based on behavioral patterns, not just past incidents
- **Fleet-level analytics.** Governance posture is assessed across all agents, all workstreams, and all time periods — not just per-agent or per-incident
- **Cross-organization learning.** Anonymized behavioral patterns from multiple organizations improve detection capabilities for all participants (data network effect)
- **Compliance reporting.** Governance data is formatted for specific regulatory frameworks and audit requirements

Risk profile: The governance system detects drift before it produces visible failures. Novel failure modes are identified through behavioral trend analysis rather than post-incident review.

Regulatory posture: Comprehensive compliance program addressing EU AI Act, NIST AI RMF, ISO 42001, and sector-specific requirements through predictive governance capabilities. Provides evidence of continuous improvement. Level 3 capabilities (fleet-level analytics, cross-organization learning, predictive risk assessment) are extrapolated from the framework's operational trajectory; they have not been demonstrated in the framework's primary operational context. Organizations should conduct independent legal review before relying on these mappings for compliance purposes.

Key gap: Governance parameters are adjusted manually based on trend data. The system informs decisions but does not adapt automatically.

Level 4: Adaptive Governance (Standard-Setting)

State: Governance mechanisms adapt based on detected drift patterns and empirical performance data. The organization contributes to the governance standard that others measure against.

Capabilities introduced:

- **Adaptive calibration.** Governance parameters (oversight levels, gate stringency, autonomy boundaries) adjust based on empirical data, within operator-approved policies
- **Governance certification.** The organization's governance program is independently assessable against published standards
- **Standard contribution.** The organization participates in governance standard development, contributing patterns, methodologies, and operational data to the broader community
- **Continuous methodology evolution.** The governance methodology itself is a versioned, improving artifact — not a static policy document

Risk profile: Minimal. The governance system learns from every interaction and improves continuously. Novel failure modes are detected, classified, and addressed through the incident-driven learning loop.

Regulatory posture: The organization contributes to governance standard development. When regulators develop more specific AI agent governance requirements, organizations at Level 4 are positioned to inform those requirements. Level 4 capabilities (adaptive calibration, governance certification, standard contribution) are aspirational and have not been demonstrated in the framework's primary operational context. Organizations should conduct independent legal review before relying on these mappings for compliance purposes.

---

3. Self-Assessment Framework

Organizations can assess their current maturity level by evaluating capabilities across five dimensions:

Dimension 1: Lifecycle Management

Maturity Level	Capability
Level 0	No formal lifecycle. Agents receive instructions and produce output.
Level 1	Defined phases with explicit transitions. Human approval at each gate.
Level 2	Prerequisite validation at gates. Structured impact statements for approval decisions.
Level 3	Lifecycle analytics. Phase duration, gate pass rates, and remediation cycles tracked across fleet.
Level 4	Adaptive lifecycle models. Phase requirements adjust based on agent and task characteristics.

Dimension 2: Behavioral Monitoring

Maturity Level	Capability
Level 0	No monitoring. Failures discovered when output is incorrect.
Level 1	Manual monitoring. Known failure patterns recognized by human operators.
Level 2	Automated detection of known patterns during agent execution.
Level 3	Trend analysis identifies emerging patterns before they produce incidents.
Level 4	New patterns automatically classified and integrated into monitoring.

Dimension 3: Trust Calibration

Maturity Level	Capability
Level 0	No calibration. All agents treated identically.
Level 1	Implicit calibration. Operators develop intuitions about agent reliability.
Level 2	Explicit calibration. Agent behavioral history informs oversight levels.
Level 3	Predictive calibration. Trend data identifies agents at risk of failure.
Level 4	Adaptive calibration. Autonomy boundaries adjust within policy constraints.

Dimension 4: Incident Management

Maturity Level	Capability
Level 0	No incident tracking. Failures addressed ad hoc.
Level 1	Incident log maintained. Root cause analysis performed.
Level 2	Incidents linked to behavioral patterns. Governance improvements traceable to incidents.
Level 3	Incident trends analyzed. Pattern frequency and severity tracked over time.
Level 4	Incident-driven methodology improvements deployed automatically within policy boundaries.

Dimension 5: Audit and Compliance

Maturity Level	Capability
Level 0	No audit trail. Governance decisions unrecorded.
Level 1	Basic audit trail. Phase transitions, approvals, and gate decisions recorded.
Level 2	Structured audit trail. Impact statements, validation results, and evidence captured.
Level 3	Compliance-formatted reporting. Audit data exportable for specific regulatory frameworks.
Level 4	Continuous attestation. Governance posture independently assessable against published standards.

---

4. Progression Path

From Level 0 to Level 1

The most critical transition. Moving from ungoverned to reactive governance requires:

1. **Define a lifecycle model.** Start with the basic MOIAS lifecycle: Plan, Build, Review, Approve. Customize phases for the organization’s workload types.
2. **Designate a human operator.** Someone must be responsible for phase transition approvals. This role is the governance control point.
3. **Start an incident log.** Begin documenting every deviation from expected agent behavior. This log becomes the foundation for all future governance

improvements.

4. **Learn the behavioral pattern taxonomy.** Familiarize operators with the eight documented failure patterns (patterns.md). Manual recognition of these patterns is the first detection capability.

From Level 1 to Level 2

Moving from reactive to structural governance requires:

1. **Automate behavioral detection.** Implement monitoring for the known behavioral patterns during agent execution, not just during post-incident review.
2. **Add prerequisite validation.** Phase transitions should be checked against plan requirements automatically.
3. **Implement impact statements.** Every phase transition request should include a structured risk assessment for the approving human.
4. **Begin trust calibration.** Track agent behavioral history and use it to inform oversight levels.

From Level 2 to Level 3

Moving from structural to predictive governance requires:

1. **Implement trend analysis.** Analyze validation logs, incident data, and behavioral metrics over time to identify drift trajectories.
2. **Deploy fleet-level analytics.** Assess governance posture across all agents and workstreams, not just per-agent.
3. **Enable cross-organization learning.** Participate in anonymized behavioral pattern sharing to improve detection capabilities.

From Level 3 to Level 4

Moving from predictive to adaptive governance requires:

1. **Define adaptation policies.** Specify how governance parameters may adjust based on empirical data, with human-approved boundaries.
2. **Establish certification readiness.** Prepare the governance program for independent assessment.
3. **Contribute to the standard.** Share governance patterns, methodology improvements, and operational data with the broader community.

---

5. Mapping to Regulatory Frameworks

> ***Alignment guidance only — not a compliance certification.*** *The mappings below reflect this framework's operational governance capabilities relative to regulatory requirements. Applicability of EU AI Act requirements depends on whether a system meets Annex III high-risk classification criteria — which must be assessed independently for each deployment. Organizations should not rely on these mappings as evidence of compliance without independent legal review.*

The maturity model maps to regulatory requirements as follows:

Requirement	Level 0	Level 1	Level 2	Level 3	Level 4
Human oversight (EU AI Act Art. 14)	None	Basic	Structured	Informed	Adaptive
Risk management (EU AI Act Art. 9)	None	Incident-based	Pattern-based	Predictive	Continuous
Record-keeping (EU AI Act Art. 12)	None	Basic audit trail	Structured audit	Compliance-formatted	Independently attestable
Continuous monitoring (EU AI Act Art. 9.2)	None	Manual observation	Automated detection	Trend analysis	Adaptive monitoring
NIST AI RMF GOVERN	Not addressed	Basic structures	Documented practices	Organizational integration	Standard contribution
NIST AI RMF MEASURE	Not addressed	Manual measurement	Automated measurement	Predictive measurement	Self-improving measurement
ISO 42001 PDCA cycle	Not addressed	Plan-Do	Plan-Do-Check	Full PDCA	Continuous PDCA

NIST AI RMF GOVERN subcategory coverage: The framework addresses GV-2 (accountability structures — operator authority map, separated authority), GV-5 (policies and controls — governance directives, phase gate policies), and GV-6 (oversight mechanisms — operator-as-router, behavioral monitoring, override capability). GV-1 (organizational context and risk tolerance) is partially addressed through the maturity self-assessment. GV-3 (workforce diversity and inclusion in AI risk management) is inherently organizational and not addressed by this operational

governance methodology — organizations should supplement with their own GV-3 policies. GV-4 (AI risk culture) is partially addressed through the Human Drift Observation and structured decision surfaces.

NIST AI 600-1 (Agentic AI Profile): NIST AI 600-1 (July 2024) identifies specific risk dimensions for agentic and generative AI systems. Several of these dimensions are addressed by this framework: confabulation risk (BP-001: Inference Over Execution maps directly to AI 600-1's confabulation risk dimension), autonomous task completion risks (addressed through phase-gated lifecycle and operator-as-router), and human oversight requirements for agentic systems (addressed through the five operator control points and Human Drift Observation). Organizations positioning this framework for federal contexts should develop an explicit AI 600-1 mapping for their deployment.

Federal deployment context: OMB Memorandum M-24-10 (March 2024) and EO 14110 impose additional AI governance requirements on federal agencies and federal contractors, including AI use case inventories, Chief AI Officer designation, and rights-impacting and safety-impacting AI assessments. These requirements are outside the scope of this framework and must be addressed separately by implementing organizations.

---

6. Common Assessment Findings

Based on observations from production multi-agent operations, the most common findings at each level:

Organizations at Level 0 typically discover:

- Agent output quality is inconsistent and unpredictable
- Failures are discovered by end users, not by governance processes
- No evidence exists to demonstrate governance to regulators or auditors

Organizations transitioning to Level 1 typically discover:

- Implementing phase gates immediately surfaces previously invisible failures
- The incident log reveals patterns that were being normalized
- Human operators become the primary bottleneck — and the primary safety mechanism

Organizations transitioning to Level 2 typically discover:

- Automated detection catches behavioral patterns that human operators missed
- Trust calibration reveals that agent reliability varies significantly by task type

- The governance system itself becomes a source of operational data

---

7. Limitations

This maturity model was developed in the context of a governed software production platform. Organizations in different domains, at different scales, or with different regulatory requirements may need to adapt the assessment dimensions and progression criteria. The maturity levels describe a general progression; specific capability requirements at each level may vary by operational context.

Version History

Version	Date	Author	Description
1.0.0	2026-02-26	John J. McCormick	Initial publication
2.0.0	2026-02-28	John J. McCormick	Metadata standardization; citation, version, and PDF fields moved to frontmatter

---