From Trustworthy AI to Trustworthy Operational AI:
Convergence of AI Governance and Agentic Engineering

Deepinder Sidhu
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County
sidhu@umbc.edu

Abstract

AI governance frameworks and operational agentic AI platforms have emerged from different directions. Governance frameworks begin with risk, compliance, security, and control. Operational agentic platforms begin with execution, workflow orchestration, mission management, and human-supervised action. Despite these different origins, both approaches are converging on common architectural requirements including human oversight, policy enforcement, runtime monitoring, auditability, observability, and governance. This paper examines that convergence and argues that governance is necessary but not sufficient for operational AI systems. Operational systems require additional capabilities including mission runtime, workflow conformance, digital twin validation, mission observability, and mission assurance. We distinguish trustworthy AI from trustworthy operational AI and argue that the latter requires multi-level observability and validation across agent, workflow, mission, and enterprise state spaces. The paper positions AI governance as an essential foundation while identifying what remains missing when AI systems become participants in mission execution.

1 Introduction

AI governance has rapidly become a priority for enterprises, government agencies, and technology vendors. Gartner describes AI Trust, Risk, and Security Management (AI TRiSM) as a framework and set of technical capabilities intended to ensure that AI systems are trustworthy, secure, and compliant through continuous monitoring, validation, and enforcement [1]. Gartner's 2025 Market Guide states that the AI TRiSM market comprises four layers of technical capabilities that enforce AI governance policies [2]. Gartner also describes AI governance platforms as an emerging market that provides central oversight of AI, risk-management frameworks, and execution of controls [**?**]. NIST's AI Risk Management Framework similarly emphasizes governance, mapping, measurement, and management of AI risk [3].

At the same time, practitioners building operational agentic AI systems have approached the problem from a different direction. Rather than beginning with enterprise risk, they begin with execution. They ask how agents coordinate, how workflows are controlled, how state is managed, how human approval is enforced, how policies are applied at runtime, and how mission outcomes are validated.

These two communities use different language but increasingly arrive at similar architectural requirements. Both require observability, runtime enforcement, human oversight, policy control, auditability, and governance. This convergence is significant because it suggests that a set of common design patterns is emerging for operational AI systems.

However, convergence does not mean completeness. Governance frameworks tend to focus on making AI systems trustworthy. Operational environments require a stronger condition: AI systems must be trustworthy participants in mission execution. This requires additional concepts not fully captured by generic governance

models, including mission runtime, workflow conformance, digital twin validation, mission assurance, and multi-level observability.

2 The Top-Down AI Governance Path

Top-down AI governance begins with organizational concerns: risk, compliance, security, trust, policy enforcement, and auditability. The simplified progression is:

$$\text{Risk} \rightarrow \text{Governance} \rightarrow \text{Controls} \rightarrow \text{Runtime Enforcement} \rightarrow \text{Observability} \rightarrow \text{Human Oversight}.$$

This path asks: *How can organizations control, govern, and trust AI systems?*

3 The Bottom-Up Agentic Engineering Path

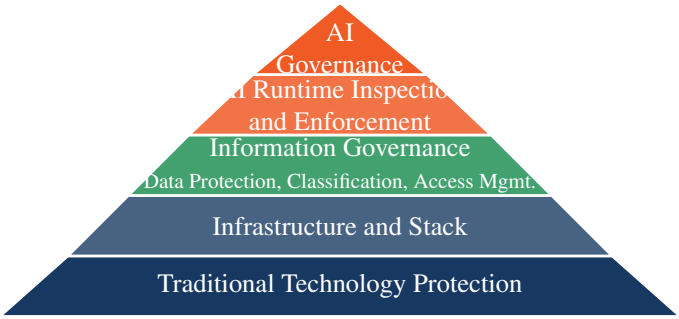
Operational agentic systems begin from execution. Developers building mission-oriented multi-agent systems encounter practical execution problems: agents must coordinate, workflows must progress, state must be managed, tools must be invoked safely, humans must approve critical actions, behavior must be observable, and failures must be detected and handled. The simplified progression is:

$$\text{Execution} \rightarrow \text{Workflow} \rightarrow \text{Mission Runtime} \rightarrow \text{Runtime Control} \rightarrow \text{Observability} \rightarrow \text{Human Oversight}.$$

This path asks: *How can operational AI systems be made reliable, governable, and scalable?*

4 Gartner AI TRiSM and the Governance Pyramid

Figure 1 summarizes the AI TRiSM technology functions as commonly presented: traditional technology protection, infrastructure and stack, information governance, AI runtime inspection and enforcement, and AI governance. The figure is redrawn as an analytical schematic rather than copied from a proprietary source.



AI TRiSM Technology Functions

Figure 1: Governance-oriented AI TRiSM technology-function pyramid, redrawn as an analytical schematic from public Gartner descriptions of AI TRiSM layers [2, 1].

The governance pyramid is valuable. It clearly identifies the technical foundations needed to secure, monitor, and govern AI systems. Its apex is AI governance. This is appropriate for the problem it is solving: how to make AI systems trustworthy, secure, and compliant.

5 The CSA Trustworthy Operational AI Pyramid

Operational AI systems require all of the AI TRiSM functions, but they also require additional layers. Figure 2 presents the CSA Trustworthy Operational AI Pyramid. The pyramid preserves the foundation of security, infrastructure, information governance, runtime enforcement, and AI governance, but extends upward into validation, multi-level observability, mission runtime, and mission assurance.

Figure 2: CSA Trustworthy Operational AI Pyramid. The model retains the governance foundations of AI TRiSM but extends upward into operational validation, multi-level observability, mission runtime, and mission assurance.

The change in apex is significant. In the governance pyramid, the apex is AI governance. In the operational pyramid, the apex is mission assurance. This does not reduce the importance of governance. It clarifies its role. Governance is necessary, but mission assurance is the operational objective.

6 Architectural Convergence and Extension

Architectural Requirement	Governance Path	Agentic Engineering Path
Human oversight	Required for accountability	Required for controlled execution
Policy enforcement	Required for compliance	Required for safe runtime behavior
Runtime monitoring	Required for risk management	Required for operational control
Auditability	Required for governance	Required for debugging and assurance
Observability	Required for inspection	Required for state reconstruction
Identity and access control	Required for security	Required for tool and data control
Workflow control	Emerging	Required
Mission runtime	Limited	Required
Digital twin validation	Limited	Required for operational assurance

Table 1: Convergence and divergence between AI governance and agentic engineering.

The convergence suggests that certain design patterns are not accidental. They appear because operational AI systems require them.

7 What Operational AI Adds

Governance frameworks correctly identify important requirements. However, when AI systems become operational participants, additional layers become necessary.

7.1 Mission Runtime

A mission runtime manages the execution of mission-oriented workflows. It tracks workflow state, agent roles, human approvals, escalation paths, policies, task status, and mission context. Generic governance frameworks may describe monitoring and enforcement, but they do not necessarily define mission execution as a first-class architectural construct.

7.2 Workflow Conformance

Operational systems require a distinction between intended workflow behavior and actual execution. A logical workflow specifies what should happen. A physical workflow records what actually happened through logs, traces, messages, API calls, or packet captures.

$$\text{Physical Workflow} \overset{?}{\cong} \text{Logical Workflow}$$

This is analogous to protocol conformance testing. The existence of traces is not enough. One must evaluate whether observed execution conforms to specification.

7.3 Digital Twin Validation

Operational AI systems should not be evaluated solely in production. Digital twins, emulation environments, cyber ranges, and representative testbeds provide controlled environments for validation. These environments allow organizations to test AI-assisted workflows, failure modes, human approval paths, and mission effects before deployment.

7.4 Mission Assurance

Mission assurance concerns the effect of AI-enabled workflows on mission outcomes. An AI agent may behave correctly and still degrade mission effectiveness. Therefore:

$$\text{Agent Correctness} \not\Rightarrow \text{Mission Success}.$$

Mission assurance requires that AI behavior be evaluated relative to mission objectives, dependencies, risks, and operational effects.

8 Multi-Level Observability

Current AI governance discussions frequently identify observability as important, but the term is often underspecified. As argued in [4], observability is not singular. Operational AI systems require observability across multiple state spaces.

$$\mathbf{O} = (O_A, O_W, O_M, O_E)$$

where O_A is agent observability, O_W is workflow observability, O_M is mission observability, and O_E is enterprise observability. These quantities correspond to different state spaces and answer different questions:

$$O_A \neq O_W \neq O_M \neq O_E, \quad O_A \to 1 \not\Rightarrow O_M \to 1.$$

Perfect visibility into an agent does not imply visibility into mission effects.

9 Trustworthy AI vs. Trustworthy Operational AI

Trustworthy AI asks whether an AI system is safe, reliable, fair, secure, governed, and compliant. These are necessary requirements. Trustworthy operational AI asks additional questions: Can the AI participate in controlled mission execution? Can workflow state be reconstructed? Can policy enforcement be verified? Can human approval paths be audited? Can mission impact be measured? Can behavior be validated in representative environments?

Thus:

Trustworthy Operational AI = Trustworthy AI+Mission Runtime+Workflow Conformance+Mission Assurance+Operatio

This equation is an architectural decomposition. It states that governance is necessary but not sufficient.

10 Implications for Government and Cybersecurity Systems

Government systems increasingly deploy AI for cybersecurity, threat hunting, knowledge management, mission planning, analytics, and operational support. In these environments, the consequences of AI behavior are not limited to model outputs. AI-enabled workflows may affect mission readiness, enterprise risk, operational continuity, and policy compliance.

Therefore, future AI-enabled government systems should include AI governance, runtime policy enforcement, human approval, workflow observability, mission observability, digital twin validation, workflow conformance testing, and mission assurance mechanisms. This moves beyond generic trustworthy AI and toward trustworthy operational AI.

11 Conclusion

AI governance frameworks and operational agentic engineering approaches are converging on common architectural requirements. This convergence is important and validates the growing emphasis on governance, observability, runtime control, and human oversight.

However, governance alone is not enough. Operational AI systems require mission runtime, workflow conformance, multi-level observability, digital twin validation, and mission assurance. The central transition is from trustworthy AI to trustworthy operational AI.

The top-down governance path correctly identifies the need for control. The bottom-up agentic engineering path shows what additional structures are required when AI systems participate in mission execution.

References

[1] Gartner. Ai governance requires more than policies. https://www.gartner.com/en/articles/ai-governance-trism, 2025. Accessed 2026-06-15.

[2] Gartner. Market guide for ai trust, risk and security management. https://www.gartner.com/en/documents/6185655, February 2025. Gartner summary page; accessed 2026-06-15.

[3] National Institute of Standards and Technology. Artificial intelligence risk management framework (ai rmf 1.0). https://www.nist.gov/itl/ai-risk-management-framework, 2023. Accessed 2026-06-15.

[4] Deepinder P. Sidhu. The observability ambiguity problem: Multi-level observability for operational ai systems, 2026. Working paper.