A Crisis of Confidence:
The Assurance Gap in Agentic AI

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

Draft — June 2026
Paper 3 of a coordinated research program on Operational AI Engineering and Agentic Science

Abstract

Artificial intelligence capability is advancing rapidly, but confidence in AI-enabled institutions is increasingly fragile. Recent incidents involving fabricated citations, unsupported claims, AI-generated legal authorities, incorrect customer-service guidance, questionable clinical recommendations, and flawed public reports demonstrate a recurring pattern across trusted institutions. These failures are often described as hallucinations, prompt failures, or human-review failures. This paper argues that such descriptions identify symptoms rather than the underlying cause. The emerging crisis of confidence in AI is not a failure of intelligence; it is a failure of assurance.

We define the *Assurance Gap* as the difference between increasing AI capability and the comparatively immature ability to establish hard guarantees about autonomous behavior. This gap becomes more consequential as AI systems evolve from AI-assisted tools, where humans usually decide before action, to AI-powered Agentic AI systems, where humans increasingly supervise during operation. The paper further introduces the *Guarantee Gap*: the difference between what an autonomous system can do and what can be guaranteed about what it will do and what it will not do. Drawing on protocol science and temporal properties, we identify four foundational guarantee classes for Agentic AI: Safety, Liveness, Recurrence, and Convergence. These properties provide a rigorous basis for moving beyond soft claims of trust toward hard guarantees supported by evidence.

The paper argues that regulation and governance are necessary but insufficient. Regulation can require assurance; it cannot create assurance. Closing the Assurance Gap requires AI Engineering and Agentic Science: disciplines capable of producing objective evidence through specification, verification, validation, testing, observability, governance, workflow conformance, mission assurance, and certification. This paper is positioned as the third contribution in a coordinated research program: observability creates visibility, trustworthy operational AI creates operational control, and assurance establishes guarantees. The future of AI will be determined not only by what autonomous systems can do, but by what can be guaranteed about what they will do and what they will not do.

Keywords: Agentic AI, Assurance Gap, AI Engineering, Agentic Science, Operational Confidence, Safety, Liveness, Recurrence, Convergence, AI Governance, Mission Assurance.

1 The Crisis of Confidence

Artificial intelligence has entered institutions whose authority depends on expertise, evidence, professionalism, and public trust. Consulting firms advise governments and enterprises. Courts and law firms establish

legal confidence. Healthcare systems influence clinical decisions. Scientific publications structure knowledge. Transportation providers operate safety-critical services. Government agencies publish information that influences public policy. These institutions are not peripheral. They are centers of societal confidence.

The most concerning AI failures are therefore not those occurring in obscure experiments or low-consequence demonstrations. They are failures appearing within institutions whose primary societal role is to create confidence. When such institutions publish unsupported claims, rely on fabricated citations, or operationalize incorrect AI-generated information, the result is not merely an isolated error. It is erosion of confidence in the institutional mechanisms by which society validates information and decisions.

The crisis of confidence in AI is not a failure of intelligence. It is a failure of assurance. We built intelligence faster than we built assurance.

1.1 Trusted Institutions and Visible Failures

Recent public incidents illustrate the pattern. A KPMG report on agentic AI was withdrawn after organizations named in the report challenged claims about their AI use; reporting and investigations indicated fabricated or inaccurate claims and citations [5, 14, 17]. Deloitte Australia agreed to provide a partial refund to the Australian government after a report produced with generative AI was found to contain errors, including non-existent academic references and a fabricated court quotation [3, 15]. EY reportedly removed a cybersecurity report after investigators identified AI-generated text, fake citations, and incorrect references [7, 16]. These examples are particularly striking because professional-services organizations sell expertise, risk management, audit, compliance, and assurance. The most visible assurance failures are increasingly occurring inside organizations whose business is assurance itself.

The same pattern appears outside consulting. In *Mata v. Avianca*, attorneys submitted legal filings containing non-existent cases generated by ChatGPT and were sanctioned by a federal court [21]. In *Moffatt v. Air Canada*, Air Canada was held liable after a chatbot provided incorrect information about bereavement fares [4, 2]. In healthcare, IBM Watson for Oncology was criticized after reports that it produced unsafe or incorrect cancer-treatment recommendations in some examples [11, 6]. In journalism, CNET issued corrections on many AI-written articles after errors were identified [19, 18]. In scientific publishing, analyses and commentary have warned that AI-generated hallucinated citations are entering the literature at significant scale [9, 10]. In government reporting, AI-linked or potentially AI-generated citation failures have raised concerns about public policy documents [20].

These are different domains. The visible failures differ. The underlying structure is similar.

1.2 What Happened and How It Happened

For this paper, the relevant question is not simply what failed, but how the failure entered an authoritative workflow. In the KPMG, Deloitte, and EY examples, inaccurate or unsupported material appeared in documents carrying institutional authority. In the legal examples, AI-generated legal authorities were incorporated into filings without sufficient verification. In the Air Canada case, AI-generated customer guidance conflicted with policy. In the healthcare example, technical capability did not automatically translate into operationally validated clinical confidence.

The common pattern is shown in Figure 1. AI produces an output. A human or institution accepts that output. An assurance layer is missing or insufficient. The output is operationalized through publication, filing, recommendation, customer interaction, or decision support. The result is institutional, business, clinical, legal, or mission impact.

The objective is not to reduce trust in institutions. The objective is to ensure that institutions possess sufficient evidence to justify that trust in the age of AI. Trust should not be transferred automatically from institutions to autonomous systems. It should be earned through assurance and supported by guarantees.

Table 1: Representative AI assurance failures across trusted institutions. The purpose is not to assign exclusive blame to AI systems, but to identify recurring assurance failures in high-trust environments.

Institution Type	Representative Example	What Happened	Assurance Failure
Professional services	KPMG	Agentic AI report withdrawn after disputed claims, inaccurate case studies, and apparent hallucinated citations.	Verification and validation failure.
Professional services	Deloitte Australia	Government report revised after false references and a fabricated court quotation were identified.	Provenance and evidence validation failure.
Professional services	EY Canada	Cybersecurity report reportedly removed after fake citations and incorrect references were identified.	Citation verification and publication assurance failure.
Legal institutions	*Mata v. Avianca*	Court filings cited non-existent legal precedents generated by AI.	Human oversight and source verification failure.
Transportation	Air Canada chatbot	Customer received incorrect fare-policy guidance from chatbot; company held liable.	Policy-conformance and governance failure.
Healthcare	Watson for Oncology	Clinical recommendations were criticized as unsafe or incorrect in reported examples.	Operational validation and clinical assurance failure.
Scientific publishing	Hallucinated citations	AI-generated or invalid references appear in scientific manuscripts and publications.	Scholarly verification and evidence-validation failure.
Government/public reports	AI-linked citation failures	Public reports contained faulty or potentially AI-generated references.	Public authority without adequate assurance.

1.3 From AI-Assisted Systems to AI-Powered Agentic AI

Most public debacles so far have occurred in AI-assisted systems. A human remained involved before the operational action: a consultant reviewed a report, a lawyer filed a brief, an editor published an article, a clinician received a recommendation, or a customer acted on chatbot guidance. The human was in the chain, even when the review was insufficient.

AI-powered Agentic AI systems change the human role. In operational agentic systems, humans may remain involved, but they increasingly supervise during operation rather than decide before every action. Agents may reason, plan, coordinate, invoke tools, communicate with other agents, and initiate actions. The question is not whether humans are present. The question is where humans sit in the control loop.

The Assurance Gap becomes more consequential as decision authority shifts from direct human decision-making to human supervision of autonomous operations. AI-assisted systems generate information. AI-powered systems generate consequences. Today's failures primarily involve information quality. Tomorrow's failures may involve decisions, actions, financial effects, safety hazards, mission degradation, and systemic risk.

1.4 Position in a Coordinated Research Program

This paper is the third contribution in a coordinated research program addressing foundational challenges in Operational AI and Agentic AI systems. The first paper, *The Observability Ambiguity Problem*, addressed

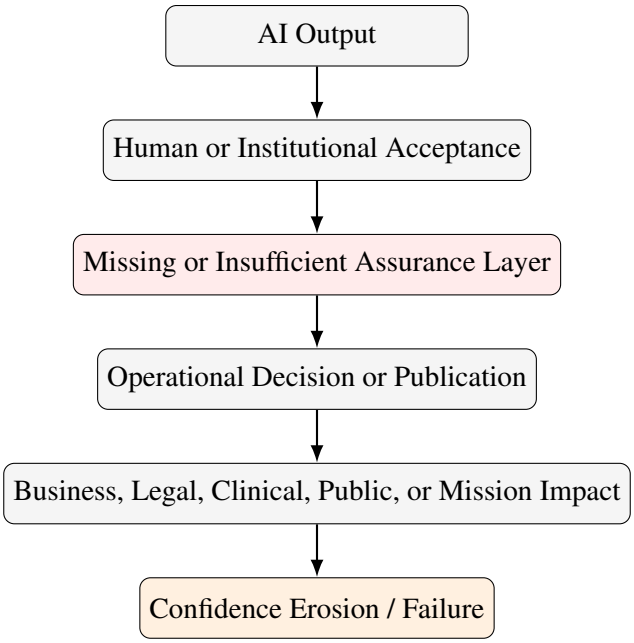

Figure 1: Anatomy of an AI assurance failure. The proximate error may appear to be a hallucination, but the operational failure often occurs because an AI output is accepted and used without sufficient assurance.

the visibility challenge: can we accurately determine what an AI system is doing? It argued that observability must specify the system, state space, abstraction level, and operational context being observed. Its core conclusion was simple: one cannot govern, validate, test, or assure what one cannot observe [13].

The second paper, *From Trustworthy AI to Trustworthy Operational AI*, addressed the operational control challenge: can AI systems be safely operated in real-world environments? It argued that governance is necessary but insufficient for operational AI. Trustworthy operation requires runtime inspection and enforcement, workflow conformance, mission runtime, digital twins, multi-level observability, and mission assurance [12].

This paper addresses the assurance challenge: can organizations establish objective guarantees regarding AI behavior? Observability creates visibility. Trustworthy Operational AI creates operational control. Assurance establishes guarantees. Together, the three papers form the progression shown in Figure 2: observe, control, and assure.

The desired outcome is not three independent papers, but a coherent trilogy addressing visibility, control, and assurance. In this structure, observability produces evidence, operational trustworthiness provides control mechanisms, and assurance transforms evidence and controls into guarantees.

2 The Assurance Gap

AI capability is advancing rapidly. Models perform better on benchmarks, agents use tools, systems retrieve external information, workflows coordinate multiple models, and organizations increasingly deploy AI in business, cyber, legal, financial, healthcare, scientific, and government processes. Yet assurance mechanisms have not advanced at the same rate.

We define the *Assurance Gap* as the difference between demonstrated AI capability and demonstrated AI assurance:

$$\text{Assurance Gap} = \text{AI Capability} - \text{AI Assurance}.$$

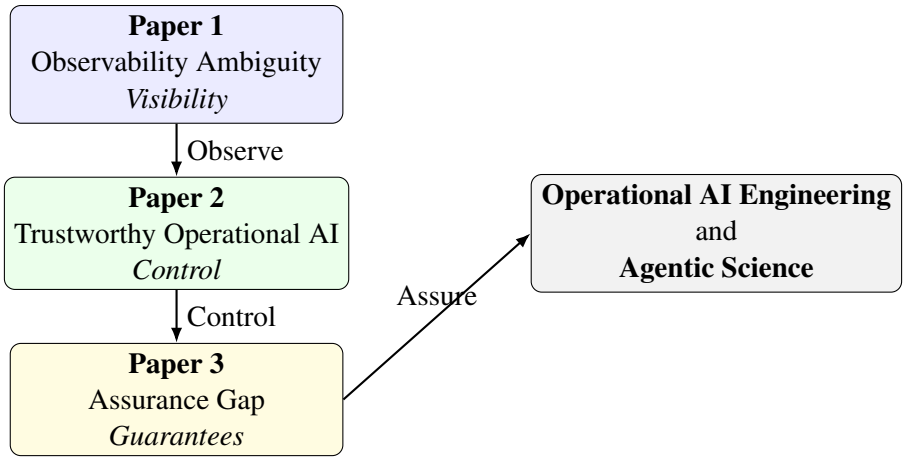

Figure 2: Coordinated research program. Paper 1 addresses visibility, Paper 2 addresses operational control, and Paper 3 addresses assurance and guarantees. Together they provide a foundation for Operational AI Engineering and Agentic Science.

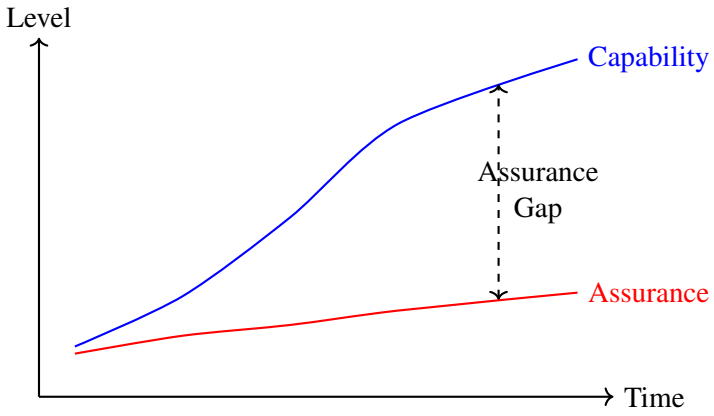

Figure 3: The Assurance Gap. AI capability is increasing faster than the ability to establish evidence, guarantees, and operational confidence regarding autonomous behavior.

This expression is not intended as a numeric equation. It is a conceptual decomposition. Capability describes what an AI system can do. Assurance describes the evidence and guarantees that establish what the system will do, what it will not do, and under what conditions its behavior remains within acceptable bounds.

2.1 Assurance and Operational Confidence

In this paper, *assurance* is the collection of evidence, analyses, tests, observations, controls, and certifications used to establish guarantees regarding the behavior of autonomous systems. Assurance is not the same as trust. Trust can be claimed. Assurance must be demonstrated. Guarantees must be established.

Operational Confidence is the degree to which stakeholders possess objective evidence that an autonomous system will operate within acceptable behavioral and operational bounds. Operational confidence is not derived solely from benchmark scores, model size, vendor reputation, or institutional authority. It is derived from evidence.

Trust in the era of AI must be earned through evidence, not inherited through reputation. Institutional trust is no longer sufficient. Operational confidence must be demonstrated.

2.2 The Guarantee Gap

The Assurance Gap ultimately manifests as a *Guarantee Gap*: the difference between the capabilities an autonomous system possesses and the guarantees that can be established regarding its behavior. Organizations increasingly know what AI can do. They often know less about what it will do, what it will not do, and which behaviors can be guaranteed under operational conditions.

The future of AI will be determined not only by what autonomous systems can do, but by what can be guaranteed about what they will do and what they will not do. Confidence can be claimed. Guarantees must be demonstrated.

2.3 Systemic Risk and Worldwide Implications

The significance of recent AI assurance failures extends beyond the individual incidents themselves. While many public failures have involved reports, recommendations, citations, or customer interactions, they should be interpreted as early indicators of a broader challenge. As AI-powered systems become increasingly integrated into financial institutions, healthcare networks, transportation systems, critical infrastructure, supply chains, and defense operations, the consequences of assurance failures may expand dramatically.

What begins as an incorrect recommendation or unsupported conclusion may evolve into operational disruption, financial loss, safety hazards, or systemic instability. The concern is therefore not limited to failures observed today. The concern is the potential impact of similar assurance failures within globally interconnected systems upon which modern society increasingly depends. The larger the system, the greater the consequence of an assurance failure. The more interconnected the system, the greater the consequence of an assurance gap.

3 Why Current AI Evaluation Is Insufficient

The AI community has become exceptionally good at measuring capability. It is comparatively immature in measuring guarantees. Standard AI evaluation emphasizes accuracy, benchmark performance, latency, cost, human preference, and task-completion rate. These are useful measures, but they do not establish whether an autonomous system satisfies safety, liveness, recurrence, or convergence guarantees.

Benchmarks measure what AI can do. Assurance measures what organizations can trust it to do.

3.1 Evaluation Is Not Assurance

Evaluation answers whether a system performed well under selected conditions. Assurance asks whether sufficient evidence exists to support hard guarantees under operational conditions. A system can score highly on benchmarks and still fail when information is incomplete, tools behave unexpectedly, policies conflict, humans delay approvals, agents disagree, or adversarial actors manipulate context.

Capability without assurance creates confidence debt. Every autonomous capability introduced without corresponding assurance widens the Assurance Gap.

3.2 Observability Is Necessary but Insufficient

Observability is essential to assurance, but observability itself must be precise. Prior work introduced the Observability Ambiguity Problem: a system cannot meaningfully claim to be observable without specifying the system, state space, and abstraction level being observed [13]. Agent observability, workflow observability, mission observability, and enterprise observability answer different questions. Observability answers what happened. Assurance answers whether it should have happened.

Table 2: Capability metrics and assurance questions. Capability metrics are necessary but insufficient for operational confidence.

Metric Type	Typical Question	Assurance Question Not Answered
Accuracy	Did the model produce the expected output on a test set?	Will unsafe actions never occur during operation?
Benchmark score	How does the model compare to other models?	Will required workflow steps eventually occur?
Latency	How quickly does the model respond?	Will monitoring, audit, and validation recur throughout execution?
Cost	How expensive is inference or operation?	Will multi-agent behavior converge to stable operational states?
Human preference	Which output do reviewers prefer?	Is the output supported by verified evidence and policy conformance?
Task completion	Was a task completed?	Was it completed through an approved and observable workflow?

You cannot assure what you cannot observe. But observation alone is not assurance. Observability creates evidence. Assurance uses evidence to establish guarantees.

3.3 Governance Is Necessary but Insufficient

Similarly, governance is necessary but insufficient. Prior work on Trustworthy Operational AI argued that generic AI governance must be extended with mission runtime, workflow conformance, digital twin validation, multi-level observability, and mission assurance for operational agentic systems [12]. Governance can define policies and controls. Assurance must determine whether those policies and controls actually hold during operation.

Agentic AI does not merely require governance. It requires assurance.

3.4 Why Observability and Governance Are Not Enough

The preceding papers in this research program addressed observability and operational trustworthiness. Observability provides visibility into system behavior. Governance provides mechanisms for controlling and constraining behavior. Both are essential. However, neither observability nor governance alone establishes guarantees.

Observability creates evidence. Governance creates controls. Assurance establishes guarantees. A highly observable system may still behave incorrectly. A highly governed system may still violate mission objectives, fail to converge, bypass intended workflows, or produce unsafe outcomes. Visibility and control are therefore necessary but insufficient conditions for operational confidence.

Operational confidence ultimately depends upon the ability to establish objective guarantees regarding autonomous behavior. Assurance is the engineering discipline through which evidence, controls, validation, testing, observability, governance, and mission assurance are transformed into demonstrable guarantees. The relationship can be summarized as:

$$\text{Observability} \rightarrow \text{Evidence}, \qquad \text{Governance} \rightarrow \text{Control}, \qquad \text{Assurance} \rightarrow \text{Guarantees}.$$

Together these capabilities provide the foundation for trustworthy operational autonomy.

Table 3: Foundational guarantee classes for Agentic AI.

Guarantee	**Temporal Form**	**Meaning**	**Agentic AI Example**
Safety	$\Box P$	A required property always holds; undesirable states never occur.	Human approval is never bypassed before protected actions.
Liveness	$\Diamond P$	A desired property eventually occurs; progress is made.	Escalation eventually occurs when uncertainty exceeds threshold.
Recurrence	$\Box\Diamond P$	A property occurs repeatedly throughout operation.	Monitoring, logging, validation, or trust refresh recurs during long-running operations.
Convergence	$\Diamond\Box P$	A property eventually stabilizes and remains true.	A multi-agent planning process eventually stabilizes on a consistent mission plan.

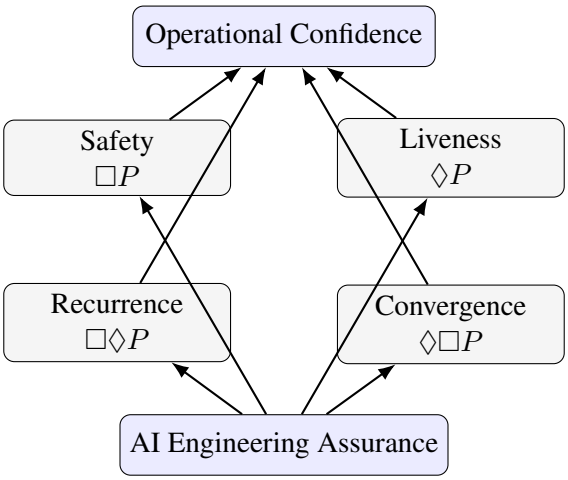

Figure 4: Foundational guarantees for Agentic AI. Assurance exists to establish evidence that safety, liveness, recurrence, and convergence properties hold within acceptable operational bounds.

4 Foundational Assurance Guarantees for Agentic AI

Assurance needs hard guarantees. Without guarantees, assurance risks becoming another term for confidence, transparency, or governance. With guarantees, assurance becomes an engineering discipline.

Drawing on protocol science, distributed systems, and temporal reasoning, we identify four foundational classes of guarantees for Agentic AI: Safety, Liveness, Recurrence, and Convergence. These are not application-specific controls. They are structural properties of autonomous behavior.

Safety, Liveness, Recurrence, and Convergence are proposed as a minimal and necessary set of foundational assurance guarantees for Agentic AI systems. Together they address four fundamental questions regarding autonomous behavior: What must never happen? What must eventually happen? What must continue to happen? What must eventually stabilize? The claim is not that these four properties are sufficient to characterize every aspect of trustworthy AI. Rather, they represent a minimal and necessary foundation upon which higher-level concepts such as governance, accountability, explainability, workflow conformance, mission assurance, and operational trustworthiness depend.

4.1 Safety: What the System Will Not Do

Safety properties characterize conditions that must always remain true. They describe what the system will not do. In Agentic AI, safety guarantees may include prohibiting unauthorized tool invocation, preventing access to protected data, blocking actions without required approval, or preventing policy violations. Safety is the foundation of guardrails.

Confidence does not come only from knowing what an AI can do. Confidence also comes from knowing what it cannot do.

4.2 Liveness: What the System Will Do

Liveness properties characterize progress. They describe what the system will eventually do. In Agentic AI, liveness guarantees may require that alerts are eventually reviewed, uncertain decisions are eventually escalated, workflows eventually reach a completion or safe-halt state, and required audit records are eventually created.

A system that is safe but never progresses may still be operationally useless. Assurance must therefore establish not only that bad things do not happen, but that required good things eventually happen.

4.3 Recurrence: What the System Continues To Do

Recurrence properties characterize repeated behavior over time. Agentic AI systems are often persistent. They monitor, reason, retrieve context, invoke tools, and interact with humans and other agents over extended periods. Assurance therefore requires guarantees that critical activities continue to recur: monitoring continues, validation continues, logs continue, policies continue to be evaluated, and human oversight opportunities continue to exist.

Recurrence is especially important for operational AI because one-time validation is insufficient for persistent autonomy.

4.4 Convergence: What the System Stabilizes To

Convergence properties characterize eventual stability. Multi-agent systems can oscillate, disagree, enter loops, or repeatedly revise plans. Convergence guarantees ask whether the system eventually reaches stable beliefs, stable plans, stable commitments, or stable mission states.

Convergence is uniquely important for AI-powered Agentic AI systems because operational confidence requires more than local agent correctness. It requires confidence that the collective behavior of agents stabilizes in ways consistent with mission objectives and constraints.

4.5 From Guarantees to Confidence

These four guarantee classes provide a formal bridge between assurance and operational confidence. Safety and liveness are well-established in protocol science and distributed systems [8, 1]. Recurrence and convergence extend the discussion toward persistent and collective agent behavior. Together they help transform assurance from a soft governance concept into a hard engineering discipline.

The defining challenge of Agentic AI is no longer capability. It is assurance. Capability determines what a system can do. Assurance determines what can be guaranteed. Operational confidence emerges when capability is constrained by demonstrable guarantees.

Table 4: Core disciplines of AI Engineering.

Discipline	Purpose	Failure if Missing
Specification	Define intended and prohibited behavior.	Wrong objectives or undefined boundaries.
Verification	Determine whether implementation satisfies specification.	Workflow defects and policy violations.
Validation	Determine whether the specified system solves the right problem.	Clinically, legally, or operationally invalid behavior.
Testing	Explore behavior under expected, adverse, and edge conditions.	Undetected defects and vulnerabilities.
Observability	Reconstruct agent, workflow, mission, and enterprise behavior.	Black-box failures and unverifiable decisions.
Human Oversight	Provide judgment, accountability, intervention, and escalation.	Automation bias and rubber-stamp approval.
Governance	Establish policies, controls, authority, and accountability.	Uncontrolled or unauthorized behavior.
Workflow Conformance	Compare physical execution with logical workflow.	Approval bypass or unauthorized sequence of actions.
Mission Assurance	Relate agent behavior to mission objectives and consequences.	Agent success with mission failure.
Certification	Establish evidence threshold for operational deployment.	Premature deployment without guarantees.

5 AI Engineering and Agentic Science

If the Assurance Gap is the problem and foundational guarantees are the objective, then AI Engineering and Agentic Science are the disciplines required to close the gap.

AI Engineering is the discipline concerned with the specification, verification, validation, testing, observability, governance, conformance, assurance, and certification of autonomous systems. Agentic Science is the scientific study of autonomous agents, their interactions, protocols, behavioral limits, guarantee properties, and operational effects.

AI Engineering begins where prompting ends.

5.1 AI Engineering Assurance Lifecycle

The AI Engineering Assurance Lifecycle identifies the minimum engineering disciplines needed to produce operational confidence. The objective of the lifecycle is to establish objective evidence that Safety, Liveness, Recurrence, and Convergence guarantees hold within acceptable operational bounds. The lifecycle is not intended as a rigid waterfall process. Rather, it identifies recurring assurance activities that must exist in some form for high-consequence AI-powered systems.

5.2 Governance of Governance

Governance is necessary but governance itself must be assured. Mature systems require governance of governance: mechanisms for determining whether governance controls are actually functioning. Are policies being enforced? Are approval workflows being bypassed? Are humans rubber-stamping decisions? Are exceptions logged? Are audits complete? Are guardrails observable?

This creates layered assurance. The agentic system is governed by policies and controls. The governance layer is itself observed, audited, and checked for conformance. Assurance mechanisms then determine

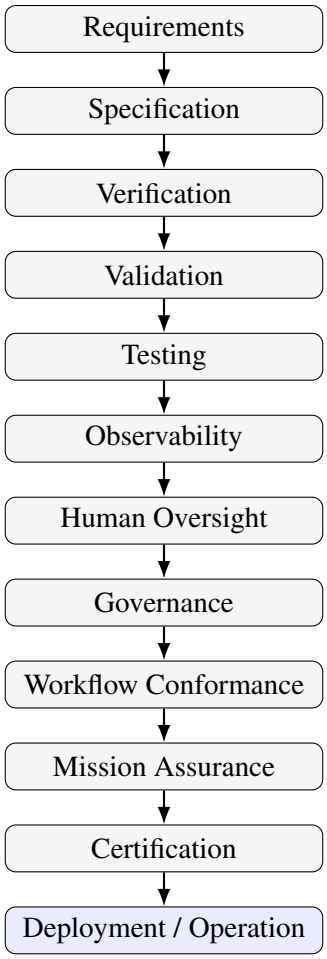

Figure 5: AI Engineering Assurance Lifecycle. Operational confidence requires evidence that spans specification, verification, validation, testing, observability, governance, conformance, mission assurance, and certification.

whether governance produces operational confidence.

5.3 Protocol Science for Agentic AI

The Internet did not become trustworthy through capability demonstrations alone. It matured through specification, protocol science, interoperability testing, conformance testing, execution-equivalent environments, packet-level observation, and operational validation. A protocol could be correctly specified and still fail in implementation. An implementation could pass unit tests and still fail during interoperability. A network could appear functional and still fail under routing dynamics, delay, loss, congestion, or adversarial behavior.

Agentic AI is approaching a similar point. Agent-to-agent interaction is not merely API compatibility. It is protocol behavior. Agents exchange messages, negotiate context, coordinate state, invoke tools, request approvals, and affect workflows. Failures may arise not only from model outputs, but from interaction protocols, state inconsistencies, missing acknowledgments, ambiguous semantics, and workflow-conformance failures.

Previous research established protocol design rules, protocol synthesis methodologies, protocol verifi-

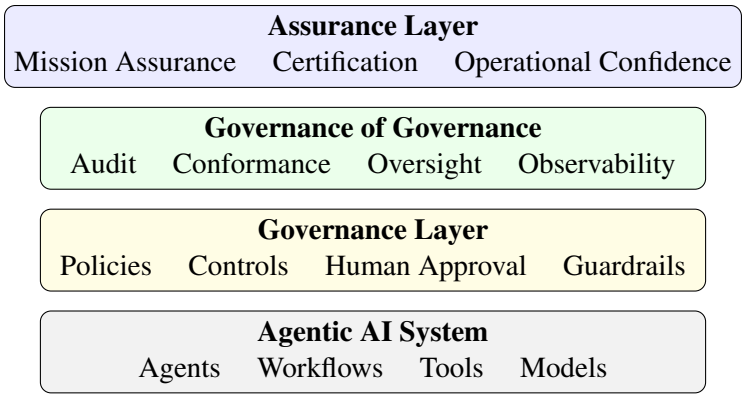

Figure 6: Layers of operational assurance. Governance alone is insufficient. Operational confidence requires governance, governance of governance, and assurance mechanisms.

cation techniques, protocol testing hierarchies, and execution-equivalent experimentation environments that transformed protocol development from an empirical activity into a scientific discipline. Similar advances are needed for Agentic AI. Agentic Protocol Science seeks to establish design rules, synthesis methodologies, verification techniques, and assurance mechanisms capable of producing correct-by-design agent interaction protocols.

This motivates Agentic Protocol Science: a discipline for specifying, analyzing, testing, and assuring agent interactions. Safety, Liveness, Recurrence, and Convergence are foundational properties for such a science.

5.4 Regulation, Oversight, and Public Safety

The growing number of visible AI failures has led many policymakers to call for stronger law, regulation, oversight, and control. This response is understandable. In domains involving healthcare, finance, transportation, critical infrastructure, national security, and government operations, autonomous decision-making can affect public safety. In high-consequence settings, assurance cannot be optional.

However, regulation alone cannot solve the underlying engineering problem. Regulation can establish requirements, accountability, transparency, audits, incident reporting, certification thresholds, and consequences for noncompliance. It cannot by itself determine whether an autonomous system satisfies safety, liveness, recurrence, and convergence guarantees.

Regulation can require assurance. It cannot create assurance. Regulation establishes accountability. Engineering establishes confidence. Public safety depends on regulation, but public confidence depends on assurance.

The long-term solution therefore requires both oversight and engineering. Regulation without engineering creates compliance without confidence. Engineering without regulation may create capability without accountability. Public safety ultimately depends on the combination of effective oversight and mature assurance disciplines capable of providing objective evidence that autonomous systems behave within acceptable operational bounds.

6 Conclusions

AI is not facing a crisis of intelligence. It is facing a crisis of confidence. That crisis is emerging because capability is advancing faster than assurance.

The most visible failures are not occurring only at the margins. They are appearing within institutions whose authority derives from expertise, evidence, professionalism, and public trust. This matters because society delegates trust to institutions. Increasingly, institutions are delegating judgment to AI. Trust in the era of AI can no longer rely solely on institutional reputation. It must be supported by evidence, assurance, and hard guarantees.

This paper introduced the Assurance Gap as the difference between rapidly increasing AI capability and comparatively immature assurance mechanisms. It further introduced the Guarantee Gap as the difference between what autonomous systems can do and what can be guaranteed regarding what they will do and what they will not do. Drawing on protocol science, we identified four foundational guarantee classes for Agentic AI: Safety, Liveness, Recurrence, and Convergence.

We propose Safety, Liveness, Recurrence, and Convergence as a minimal and necessary foundation for trustworthy and effective Agentic AI systems.

This paper also completes a three-part progression. The Observability Ambiguity Problem identified the need for visibility. Trustworthy Operational AI identified the need for operational control. This paper identifies the need for assurance and guarantees. Together, these capabilities establish a foundation for Operational AI Engineering and Agentic Science: disciplines whose purpose is to create autonomous systems that are observable, controllable, assured, and worthy of operational confidence. Observability provides visibility. Trustworthy Operational AI provides operational control. Assurance establishes guarantees. Together these capabilities form the foundation of operational confidence.

The solution is not less AI. The solution is better engineering. AI Engineering and Agentic Science must mature rapidly enough to establish hard guarantees for autonomous systems before AI-powered workflows become deeply embedded in financial, healthcare, transportation, government, critical infrastructure, and defense operations.

We built intelligence faster than we built assurance. Regulation can require assurance; it cannot create assurance. The future of AI will be determined not only by what autonomous systems can do, but by what can be guaranteed about what they will do and what they will not do.

References

[1] Bowen Alpern and Fred B. Schneider. Defining liveness. *Information Processing Letters*, 21(4):181–185, 1985.

[2] American Bar Association. Bc tribunal confirms companies remain liable for information provided by ai chatbot, February 2024. Accessed June 2026.

[3] Associated Press. Deloitte to partially refund australian government for report with apparent ai-generated errors, October 2025. Accessed June 2026.

[4] British Columbia Civil Resolution Tribunal. Moffatt v. air canada, 2024 bccrt 149, 2024. Accessed June 2026.

[5] Financial Times. Kpmg report contained ai hallucinations on benefits of ai, June 2026. Accessed June 2026.

[6] Healthcare Dive. Ibm's watson gave unsafe and incorrect cancer treatment advice, July 2018. Accessed June 2026.

[7] Information Age / Australian Computer Society. Ey retracts cyber report littered with ai errors, May 2026. Accessed June 2026.

[8] Leslie Lamport. Proving the correctness of multiprocess programs. *IEEE Transactions on Software Engineering*, SE-3(2):125–143, 1977.

[9] Miryam Naddaf and Elizabeth Quill. Hallucinated citations are polluting the scientific literature. what can be done?, April 2026. Accessed June 2026.

[10] David B. Resnik. Hallucinated citations produced by generative artificial intelligence may constitute research misconduct when citations function as data in scholarly papers. *Accountability in Research*, 2026.

[11] Casey Ross and Ike Swetlitz. Ibm's watson recommended unsafe and incorrect cancer treatments, internal documents show, July 2018. Accessed June 2026.

[12] Deepinder Sidhu. From trustworthy ai to trustworthy operational ai: Convergence of ai governance and agentic engineering. Working paper, 2026.

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

[14] TechCrunch. Kpmg pulls report on ai usage due to apparent hallucinations, June 2026. Accessed June 2026.

[15] The Guardian. Deloitte to pay money back to albanese government after using ai in report, October 2025. Accessed June 2026.

[16] The Indian Express. Ey withdraws report over ai hallucination errors, fake data and citations, May 2026. Accessed June 2026.

[17] The Register. Kpmg's ai report turns into a demo of ai hallucinations, June 2026. Accessed June 2026.

[18] The Verge. Cnet found errors in more than half of its ai-written stories, January 2023. Accessed June 2026.

[19] The Washington Post. Cnet used ai to write articles. it was a journalistic disaster, January 2023. Accessed June 2026.

[20] The Washington Post. White house maha report may have garbled science by using ai, experts say, May 2025. Accessed June 2026.

[21] United States District Court, Southern District of New York. Mata v. avianca, inc., 678 f. supp. 3d 443, 2023. Decision dated June 22, 2023; accessed June 2026.