



RESPONSIBLE
AI TRUST

SEPTEMBER 2026

AUDIT-READY, ALWAYS-READY

Why Continuous Assurance Must
Follow Machine-Speed Defense

PRESENTED & LED BY
LEHAR GUPTA, FOUNDER & CEO,
RESPONSIBLE AI TRUST

AUTHORS

Lehar Gupta, Founder & CEO, Responsible AI Trust
Amy Jan, Chief AI Architect | U.S. Federal Government
Lena Smart, Former CISO, MongoDB



Responsible AI Trust
Flagship Brief
#3

Powered by



MetricsLM



Silver Partner

TABLE OF CONTENTS



Foreword from the Founder

About this Brief	1
Leader's Dashboard	2
Executive Summary - Audit-Ready, Always-Ready	4
The Assurance Gap	5
The Problem - Point-in-Time Certification	6
Continuous Assurance Defined - From Fragmentation to Foresight	7
The Future State	9
Core Trust Systems - From Compliance Programs to Trust Systems	10
When Boundaries Change - Are We Still Enforcing the Right Boundary?	11
The Closed Loop of Continuous Assurance	12
Operating Model - The Closed Loop of Continuous Assurance	13
The Next Evolution - AI Continuous Assurance	14

Powered by



TABLE OF CONTENTS



Conclusion - From Point-in-Time Trust to Continuous Trust	15
What This Means for You (Now vs Next)	17
Final Words	18
Author Perspectives	
Appendix - 1, 2, 3, 4	

Powered by



MetricsLM

IBM
Silver Partner



FOREWORD FROM THE FOUNDER



At Responsible AI Trust, our purpose is simple yet vital: to strengthen global confidence in intelligent systems.

Across industries and borders, AI is redefining how decisions are made, risks are managed, and accountability is shared. With this transformation comes a collective responsibility to ensure that technology remains aligned with human values, transparency, and fairness.

Each brief we publish represents collaborative work from researchers, advisors, and practitioners who believe that trust must be earned through clarity, governance, and evidence. Together, we translate complex regulation into actionable insight, helping leaders navigate uncertainty with structure and foresight.

Responsible AI is not a trend; it is the foundation of sustainable innovation. As systems grow more capable, our frameworks for oversight must grow equally intelligent, adaptive, and globally connected.

Thank you for being part of this effort to turn principles into practice, and ideas into accountability.

Lehar Gupta

Founder & CEO, Responsible AI Trust

✉ Lehar@ResponsibleAITrust.com



ABOUT THIS BRIEF

Audit-Ready, Always-Ready

Audit-Ready, Always-Ready examines a growing enterprise challenge:

AI systems and agents are changing faster than traditional assurance can keep pace. As models, prompts, data, permissions, tools and interactions continuously change, this brief proposes a shift from point-in-time to continuous assurance.

Purpose and Motivation:

Machine-speed defense requires machine-speed assurance. This brief examines how organizations can continuously verify trust as AI systems evolve.

Audience:

For CISOs, AI governance, assurance, security, risk and compliance leaders responsible for deploying AI safely at enterprise scale.

Methodology:

Synthesizes emerging research across AI security, governance and agentic systems to propose a closed-loop model:

Define → Observe → Evaluate → Enforce → Verify → Reassess.

Version:

Public Release 1.0 - September 2026

Flagship Series: Responsible AI Trust Brief #3

Future updates will incorporate emerging standards, treaties, and evaluation benchmarks to maintain a live, global reference.

● [View the latest live version](#)

Disclaimer:

"Audit-Ready, Always-Ready" has been independently developed by Responsible AI Trust. The analyses, views, and recommendations expressed herein are solely those of the report authors and do not necessarily represent the opinions or endorsements of any organizations referenced. This publication is provided "as is," without warranty of any kind. Responsible AI Trust accepts no liability for actions taken based on its content. This document is for informational purposes only and should not be relied upon as legal or regulatory advice.

All rights reserved. No part of this report may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, without the prior written permission of Responsible AI Trust, except in the case of brief quotations used in reviews or scholarly works with proper citation. Any unauthorized sale, alteration, or redistribution of this report is strictly prohibited and should be reported to info@responsibleaitrust.com.

LEADER'S DASHBOARD



General Information

Brief Title	: Audit-Ready, Always-Ready - Responsible AI Trust
Focus	: Continuous Assurance for AI Systems & Agents
Authors	: Lehar Gupta Aryn Jan Lena Smart
Date Prepared	: September 2026

Executive Summary

AI systems are changing faster than traditional assurance can keep pace. Point-in-time assessments show what was trusted; continuous assurance determines whether that trust still holds today. The emerging requirement is a closed loop connecting governance intent, runtime behavior, enforcement, evidence and reassessment.

Strategic Insight

The assurance question is changing:

- Yesterday: Were the controls effective when we assessed them?
- Today: Are the controls effective right now?
- Next: Are we still enforcing the right boundaries?

As AI becomes more autonomous, monitoring alone is insufficient. Assurance must move closer to execution while continuously verifying whether trusted assumptions remain valid.

Executive Priorities - "Now vs Next"

Now	Next
Know what is operating: agents, models, tools, data and permissions.	Continuously observe change across the operating environment.
Define trusted boundaries from policy, risk appetite and acceptable use.	Evaluate and enforce those boundaries at machine speed.
Validate controls and maintain evidence of effectiveness.	Trigger reassessment when change invalidates trusted assumptions.

LEADER'S DASHBOARD



Continuous Assurance Model

Stage	Leadership Question	Outcome
1. Define	What should be trusted?	Boundaries
2. Observe	What is happening?	Visibility
3. Evaluate	Should this action happen?	Decision
4. Enforce	What action should we take?	Control
5. Verify	Did the control work?	Evidence
6. Reassess	Is the boundary still sufficient?	Updated trust

Key Takeaways for Leaders

- Point-in-time trust expires as systems change.
- Monitoring tells you what happened; assurance asks whether it should happen.
- Controls need machine-speed evaluation, enforcement and evidence.
- A working control does not guarantee the boundary itself remains sufficient.

Strategic Call to Action

Ask of every consequential AI system:

- What is it allowed to do?
- How do we know its controls still work?
- What happens when something materially changes?

The goal is not more governance. It is continuous confidence that AI remains within trusted operating boundaries.

Presented by

Lehar Gupta

Founder & CEO, Responsible AI Trust
Lead Author, Audit-Ready, Always-Ready
Lehar@ResponsibleAITrust.com



EXECUTIVE SUMMARY

Audit-Ready, Always-Ready

The AI industry has largely focused on accelerating detection, response, and remediation to match machine-speed threats. The next challenge is assurance.

As AI systems, AI agents, and autonomous workflows become increasingly capable, traditional assurance models are struggling to keep pace. Annual audits, point-in-time certifications, and periodic assessments were designed for environments where systems changed slowly. Modern AI environments evolve daily through model updates, prompt changes, data modifications, agent configuration changes, tool integrations, and infrastructure updates.

- A control verified and certified six months ago may no longer be effective today.
- A certification achieved last quarter does not guarantee compliance this quarter.
- An AI agent approved yesterday may behave differently tomorrow.
- The industry requires a shift from point-in-time assurance to continuous assurance.
- Organizations must become audit-ready, always-ready.



THE ASSURANCE GAP

Most organizations currently operate using a familiar cycle:

Prepare for audit
Collect evidence
Demonstrate compliance
Receive certification or attestation
Repeat next year

This model worked for traditional systems because the environment remained relatively stable between audits.

AI changes that assumption.

AI systems introduce:

Continuous model evolution
Dynamic agent behaviors
Rapid deployment cycles
Frequent third-party dependencies
Expanding regulatory requirements
Autonomous decision-making

The result is a growing gap between:

What was verified and What is actually happening today

This is the assurance gap.



THE PROBLEM

Point-in-Time Certification

Most assurance frameworks answer a historical question:

"Were controls operating effectively during the assessment period?"

Enterprise buyers increasingly need a different answer:

"Are controls operating effectively right now?"

This distinction becomes critical in AI environments where:

- New agents can be deployed in minutes or hours
- New tools can be connected instantly
- Models can be updated frequently
- New data sources appear continuously

The pace of operational change has exceeded the pace of traditional assurance.



CONTINUOUS ASSURANCE DEFINED

From Fragmentation to Foresight

Continuous assurance is the ongoing collection, validation, monitoring, and verification of trust signals across AI systems, agents, controls, and operational processes.

Rather than relying on periodic snapshots, continuous assurance creates a living representation of organizational trustworthiness.

Key characteristics include:

Continuous Evidence Collection

Evidence is generated automatically from systems rather than assembled manually before an audit.

Continuous Control Validation

Controls are tested regularly to ensure they remain effective.

Continuous Monitoring: What is happening?

Operational signals are continuously evaluated for drift, degradation, and emerging risk.

Assurance at the Point of Action: Should this action happen?

Continuous monitoring tells an organization what its AI systems and agents are doing. But as systems become more autonomous, observation alone is not enough.

The next question is not simply:

"Is this agent behaving abnormally?"

It is:

"Is this agent still allowed to do what it is trying to do, under these conditions, right now?"

An agent may have legitimate credentials, approved tools and valid permissions, yet still create risk because the environment has changed, permissions interact in unexpected ways, or another agent alters the context in which an action occurs. This shifts assurance closer to the execution path. For consequential actions, organizations increasingly need the ability to evaluate an action against current policy, identity, permissions, data sensitivity and environmental context before execution, and to allow, deny, constrain or escalate it accordingly.



CONTINUOUS ASSURANCE DEFINED

From Fragmentation to Foresight

Continuous assurance therefore evolves from observing controls to participating in the control loop.

Executable Controls: Can the decision be consistently enforced at machine speed?

As agent populations grow, this problem becomes one of scale. Enterprises may eventually operate thousands of heterogeneous agents across different models, tools, environments and business processes. Their permissions, objectives and operating context will continuously change.

Human-readable policies alone cannot operate at that speed. The emerging requirement is to translate organizational policy and risk boundaries into machine-evaluable rules that can be applied consistently at runtime.

The progression becomes:

Policy → Machine-readable rule → Runtime decision → Enforcement → Evidence

This introduces another assurance requirement: decision consistency. For high-consequence actions, organizations need confidence that the same machine-enforceable policy evaluated against materially equivalent conditions produces a predictable and reproducible outcome. Emerging approaches such as policy autoformalization offer one potential path for translating natural-language requirements into machine-enforceable policies at scale.

Continuous Traceability

Changes to controls, models, agents, and policies are tracked over time.

Continuous Trust Provenance

Organizations maintain a verifiable history of why trust decisions were made and which evidence supported them.



THE FUTURE STATE

Audit-Ready, Always-Ready

The goal is not to eliminate audits.

The goal is to make audits a natural by-product of operational excellence.

In an audit-ready, always-ready environment:

- Evidence already exists
- Controls are already validated
- Risks are already monitored
- Framework mappings are already maintained
- Assurance artifacts are continuously updated

The audit becomes verification rather than preparation.



CORE TRUST SYSTEMS

From Compliance Programs to Trust Systems

Historically, compliance programs have focused on satisfying requirements.

The future belongs to organizations that build trust systems.

Trust systems continuously answer:

- What controls exist?
- Are they working?
- What evidence supports them?
- When were they last validated?
- What changed?
- What risks remain?

These answers should be available at any moment, not only during audits.



WHEN BOUNDARIES CHANGE

Are We Still Enforcing the Right Boundary?

Runtime enforcement assumes that the operating boundary being enforced remains the right one. Agentic systems make that assumption increasingly difficult to sustain.

Unsafe outcomes do not necessarily begin with a malicious prompt or an obvious policy violation. Research on agentic systems has shown how behavior can change as environmental conditions change, and how interactions between autonomous agents can produce outcomes that are difficult to anticipate from the behavior of any individual agent alone.

Jha et al. demonstrate how routine environmental failures can lead agents toward unsafe workarounds without adversarial prompting [1]. Anthropic's research on emerging multiagent systems examines a related challenge at the system level: individually benign agent behaviors can interact in ways that produce unexpected collective outcomes [2]. Research on agentic misalignment [3] and sleeper agents [4] further illustrates how risky behavior may remain latent or emerge under particular conditions.

This creates a deeper assurance question.

It is no longer sufficient to ask:

"Did the agent remain within its operating boundary?"

Organizations must also ask:

"Is the operating boundary itself still sufficient?"

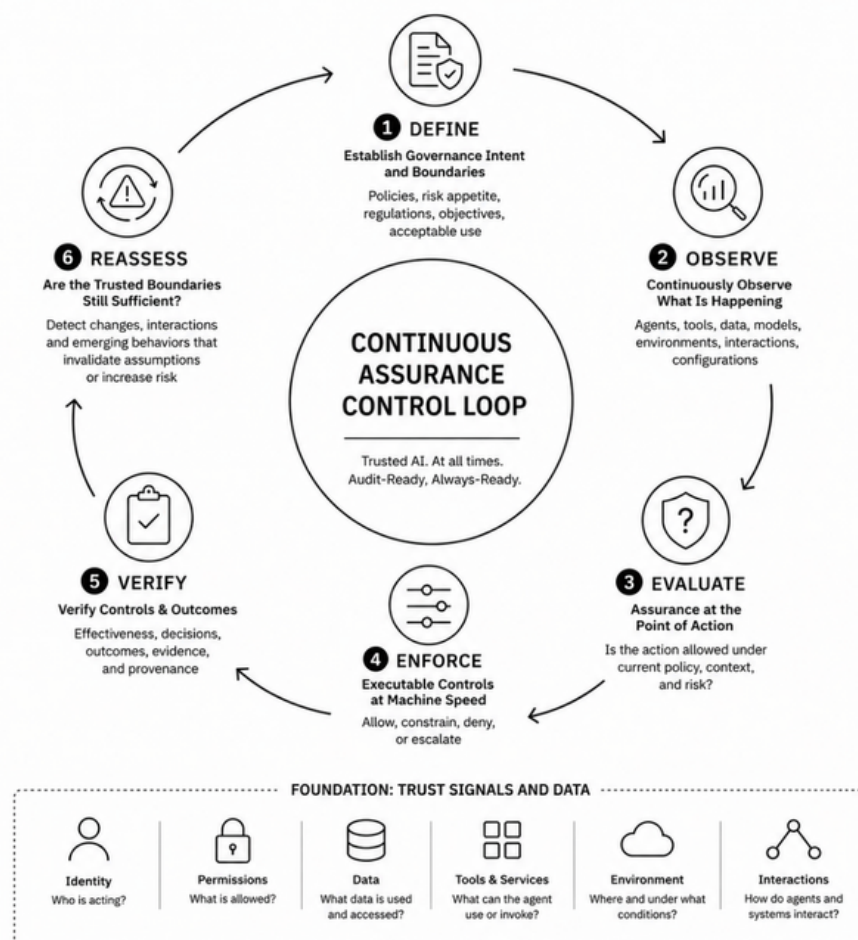
A control may operate exactly as designed while changes in the agent, its permissions, tools, environment or interactions invalidate the assumptions on which that control was established. Continuous assurance must therefore verify controls and continuously reassess whether the boundaries those controls enforce remain appropriate as the system evolves.



THE CLOSED LOOP OF CONTINUOUS ASSURANCE

AIUC-1's Defending at Machine Speed After Mythos [5] describes the need for security capabilities that can operate at machine speed. The assurance challenge that follows is whether organizations can verify, at comparable speed, that those controls and the assumptions behind them remain valid.

The next step is closing the loop between what the organization intends, what the system is permitted to do, what actually happens, and what the organization learns from it.





OPERATING MODEL

The Closed Loop of Continuous Assurance





THE NEXT EVOLUTION

AI Continuous Assurance

Machine-speed threats require machine-speed defense.

Machine-speed defense requires machine-speed assurance.

But machine-speed assurance cannot rely on monitoring alone. As autonomy increases, assurance must move closer to execution: continuously determining whether agents remain within trusted operating boundaries, enforcing those boundaries where appropriate, and reassessing them as systems and environments change.

Organizations that embrace continuous assurance will:

- Increase customer trust
- Accelerate enterprise procurement
- Respond faster to emerging risks
- Reduce audit preparation effort
- Improve regulatory readiness

Most importantly, they will maintain confidence that their AI systems remain trustworthy as they evolve.

The future is not point-in-time trust.

The future is continuous trust.

The future is audit-ready, always-ready.



CONCLUSION

From Point-in-Time Trust to Continuous Trust

AI systems are beginning to operate at a speed that traditional assurance was never designed to match. Models change, agents gain new capabilities, tools and permissions evolve, and interactions create conditions that were not present when a system was first assessed.

The challenge is therefore no longer simply proving that an AI system was trustworthy at a point in time. It is maintaining confidence that it remains trustworthy as it changes.

Assurance must become a continuous control loop

Continuous assurance connects what an organization intends, what an AI system is permitted to do, what actually happens, and what the organization learns from it.

This requires a closed loop:

Define → Observe → Evaluate → Enforce → Verify → Reassess

Monitoring provides visibility. Enforcement keeps systems within defined boundaries. Verification provides evidence that controls operated as intended. But assurance must go further: it must detect when changes in agents, permissions, tools, data, environments or interactions invalidate the assumptions behind those boundaries.

The boundary itself must be reassessed

A control can operate exactly as designed and still become insufficient.

As AI systems become more autonomous, organizations must continuously ask not only "Did the control work?" but "Are we still enforcing the right boundary?"

That distinction moves assurance from a periodic compliance activity to an operational capability.



CONCLUSION

From Point-in-Time Trust to Continuous Trust

Audit-ready becomes the outcome, not the exercise

The objective is not to eliminate audits, governance or human oversight. It is to make evidence, control validation, traceability and reassessment part of everyday AI operations.

When assurance operates continuously, an audit becomes verification rather than preparation.

Machine-speed threats require machine-speed defense.

Machine-speed defense requires machine-speed assurance.

The future is not point-in-time trust.

The future is continuous trust.

The future is audit-ready, always-ready.



WHAT THIS MEANS FOR YOU

(Now vs Next)

For **leaders**, the immediate priority is to know what AI systems and agents are operating, what they are permitted to do, and which controls are meant to keep them within trusted boundaries. Assurance should become part of the AI lifecycle, not an exercise performed after deployment.

For **security, risk and assurance teams**, move from periodic evidence collection toward continuous verification. Connect changes in models, prompts, permissions, tools, data and environments to control testing, evidence and reassessment.

For AI and product teams, design for assurance from the beginning. Consequential actions should be observable, evaluable and, where appropriate, enforceable, with clear evidence of what happened and why.

What comes next is a shift from monitoring systems to closing the assurance loop:

Define → Observe → Evaluate → Enforce → Verify → Reassess

The objective is not more governance. It is maintaining confidence that AI remains within trusted operating boundaries as it evolves.



FINAL WORDS

AI assurance faces a fundamental mismatch: AI operates continuously, while assurance is still largely periodic.

Closing that gap requires more than monitoring. Organizations need to connect governance intent to runtime behavior, verify that controls continue to work, preserve evidence of consequential decisions, and reassess trusted boundaries when the conditions around them change.

A control can operate exactly as designed while the assumptions behind it become outdated. That is why the defining question of continuous assurance is not only:

"Did the control work?"

It is:

"Are we still enforcing the right boundary?"

Machine-speed threats require machine-speed defense.

Machine-speed defense requires machine-speed assurance.

The future is not point-in-time trust.

The future is continuous trust.

The future is audit-ready, always-ready.



THE CISO PERSPECTIVE

The hardest boundary to reassess is the one the organization already trusts.



Lena Smart

former CISO, MongoDB | Tradeweb | NYPA

Board Member American Society for AI

I spent over two decades running security programs where the audit calendar was, frankly, the enemy of real assurance. This paper puts language to a frustration I lived with for years: the gap between "we passed our SOC 2" and "are we actually safe today" was always wider than boards wanted to admit, and AI has just made that gap impossible to ignore.

What strikes me most is the shift from asking whether an agent stayed within its boundary to asking whether the boundary itself is still the right one. In my CISO years, we obsessed over control effectiveness and rarely questioned control relevance - this paper correctly identifies that agentic systems demand both, continuously.

I also appreciate that the paper doesn't treat this as a purely technical problem. The move from "compliance programs" to "trust systems" is really a governance and culture shift, not just a tooling upgrade, and it's the argument I've made to boards for years about security maturity generally - just now with much less time to react.

If I have one addition from the field: the closed-loop model (Define, Observe, Evaluate, Enforce, Verify, Reassess) is sound, but enterprises will struggle most at the "Reassess" stage, because reassessing boundaries requires organizational humility that audit-driven cultures rarely reward. That's the leadership challenge underneath the technical one.

The paper's closing line - that the future is continuous trust - is the right frame, and it's one every board and CISO should internalize before their next AI deployment cycle, not after.



THE AGENTIC SYSTEMS PERSPECTIVE

When machines act on our behalf, permission is only the beginning.



Aryn Jan

Chief AI Architect | U.S. Federal Government

Founder of AJ Emtech LLC | Founding Member of AIUC-1

Member of ASFAI

My work across AI strategy, architecture, mission systems, and enterprise transformation has taught me a simple lesson: security and assurance cannot sit outside the system.

They have to be part of how the system operates. This becomes even more important as AI moves from tools that assist people to agents that can make decisions, take actions, and interact with other agents. Unlike traditional software, agents do not simply follow a fixed set of instructions. They observe, reason, make decisions, use tools, and take actions to accomplish a goal. That is what makes them powerful, but it also creates new risks. An agent can access restricted data, make an unauthorized change, or go beyond the authority we intended to give it. In the traditional approach, we prepare for an audit, collect evidence, and demonstrate that controls worked at a point in time. But that model does not fit this environment. By the time the audit is complete, the system may have changed. That is the assurance gap. Threats now move at machine speed. Our defenses, and our ability to know whether those defenses are working, have to move at the same speed. Evidence, control validation, risk signals, and traceability should not be something we assemble when someone asks for them. They should already exist as part of how the system operates.

There is another gap that becomes more important with agents: the gap between authorization and intent. Traditional access controls were designed around human users and software that generally behaved in predictable ways. They answer simple questions: Who are you? What can you access? What are you allowed to do? Agents make this harder. They plan, reason, delegate, and adapt during a task. A single request from a user can pass through several agents, tools, and systems. At each step, something can change.



THE AGENTIC SYSTEMS PERSPECTIVE

When machines act on our behalf, permission is only the beginning.



Aryn Jan

Chief AI Architect | U.S. Federal Government

Founder of AJ Emtech LLC | Founding Member of AIUC-1

Member of ASFAI

The permissions may still be valid, but the original intent can get lost. An action can be completely authorized and still be wrong. A user may ask an agent to retrieve information, for example, and a downstream agent may decide to modify the underlying data. That action may be allowed by the existing policy, but it was never what the user asked the system to do. This raises a question that traditional security controls do not really answer: Why is this action being taken, and is it still consistent with what the user intended? When machines can act on our behalf, permission becomes the beginning of the question, not the end of it.

For me, Audit-Ready, Always-Ready means moving from periodic assurance to continuous assurance. We should know what changed, what is happening now, whether our controls are still working, whether the assumptions behind those controls are still valid, and whether the actions being taken are still within the intent and authority that started the task.

We also cannot only monitor the behavior we expect. We need to understand what happens when an agent does something unexpected, when several agents interact and produce an outcome that none of them was intended to produce, or when a behavior only appears under a particular condition. This is where security has to move from static control to adaptive and dynamic control. The goal is not to create more processes or add another layer of friction. It is to make security and assurance part of the system itself.

The bottom line: if our systems can operate, adapt, and take actions at machine speed, our security and assurance must be able to see what is happening, understand why it is happening, and respond at the same speed and scale.



THE ASSURANCE PERSPECTIVE

Trust is not a state. It must be continuously verified.



Lehar Gupta

Founder & CEO, Responsible AI Trust

✉ Lehar@ResponsibleAITrust.com

Lead Author, Audit-Ready, Always-Ready

I started with a simple question: if an AI system was trusted when we assessed it, how do we know it is still trustworthy today?

That question became harder as AI moved from models we could evaluate at a point in time to agents that can use tools, access data, make decisions and act. The system we approved yesterday may not be the system operating today. A model changes. A permission expands. A new tool is connected. An interaction creates a condition we did not anticipate. Yet the evidence we rely on can still describe a version of the system that no longer exists.

That is the assurance gap.

My view is that closing it requires more than performing the same assurance activities more frequently. Assurance itself has to become part of the operating model of AI.

That is why this paper proposes a continuous loop: Define, Observe, Evaluate, Enforce, Verify, Reassess. It connects what an organization intended, what the system is permitted to do, what actually happens, whether the controls worked, and what we learn as the system changes.

For me, the most important step is the last one: Reassess.

We tend to think of assurance as proving that a control is working. But a control can operate exactly as designed while the assumptions behind it become wrong. The agent may have changed. Its permissions may have changed. Its environment may have changed. The interaction between systems may have created a risk that did not exist when the boundary was approved.



THE ASSURANCE PERSPECTIVE

Trust is not a state. It must be continuously verified.



Lehar Gupta

Founder & CEO, Responsible AI Trust

✉ Lehar@ResponsibleAITrust.com

Lead Author, Audit-Ready, Always-Ready

So the defining question is no longer only:

“Did the control work?”

It is:

“Are we still enforcing the right boundary?”

That distinction matters because continuous assurance should not become another layer of compliance that slows AI adoption. Done well, it should do the opposite. When evidence already exists, controls are continuously verified, and material change triggers reassessment, organizations can make decisions with greater confidence and respond before the next audit tells them something changed months ago.

I believe this is the shift ahead of us: from proving trust periodically to maintaining trust continuously.

The audit should become a consequence of good assurance, not the moment assurance begins.

AI will not wait for the next audit cycle. Assurance cannot either.



APPENDIX

Appendix 1 - Continuous Assurance Operating Model

Stage	Core Question	Illustrative Evidence
Define	What should the system be trusted to do?	Purpose, policy, risk appetite, approved permissions
Observe	What is happening?	Runtime events, identity, permissions, tool and data access
Evaluate	Should this action happen?	Policy decision, context, risk signals
Enforce	What should happen next?	Allow, constrain, deny, sandbox, escalate
Verify	Did the control operate as intended?	Decision record, control result, evidence trail
Reassess	Is the approved boundary still sufficient?	Material changes, new interactions, emerging behavior



APPENDIX

Appendix 2 - Illustrative Reassessment Triggers

Reassessment may be appropriate when a material change could invalidate an existing assurance decision or the assumptions on which a trusted operating boundary was established.

- Model or agent configuration
- System prompt or workflow
- Permissions or identity
- Connected tools or services
- Accessible data or data sensitivity
- Autonomy or delegated authority
- Human-approval requirements
- Operating environment
- Agent-to-agent interactions
- Control effectiveness or underlying assumptions

These triggers are illustrative rather than universal requirements. Materiality should be determined according to the system, risk context and organizational policy.



APPENDIX

Appendix 3 - Key Terms

Term	Definition
Continuous Monitoring	ongoing observation of system behavior and operational signals.
Continuous Assurance	ongoing verification that controls, evidence and trusted assumptions remain valid as the system changes.
Operating Boundary	the conditions within which an AI system or agent is intended and authorized to operate.
Material Change	a change significant enough to potentially invalidate an existing assurance decision or trusted assumption.
Trust Provenance	the evidence and decision history supporting why a system or action was considered trustworthy.



APPENDIX

Appendix 4 - References

- [1] Jha, R., Triedman, H., Bhattacharya, A. & Shmatikov, V. (2026). *Agent Meltdowns: The Road to Hell Is Paved with Helpful Agents*. arXiv:2605.19149.
- [2] Anthropic (2026). *Patterns and Problems in Emerging Multiagent Systems*. 13 August 2026.
- [3] Anthropic (2025). *Agentic Misalignment: How LLMs Could Be Insider Threats*.
- [4] Hubinger, E., et al. (2024). *Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training*. arXiv:2401.05566.
- [5] AIUC-1 Consortium (2026). *Defending at Machine Speed After Mythos*. Executive briefing for CISOs and security leaders, 5 June 2026.



RESPONSIBLE
AI TRUST

SEPTEMBER 2026

BUILDING TRUST IN INTELLIGENT SYSTEMS BEGINS WITH YOU.

At Responsible AI Trust, we believe alignment is not a debate; it's a design principle. Every framework, every audit, every standard leads to one shared goal: verifiable trust. Whether you're an enterprise **leader** shaping governance strategy, a **researcher** mapping global standards, or a **sponsor** advancing responsible innovation, your participation helps define how the world governs AI.

Join us in building the foundation for measurable, interoperable, and accountable AI.

✉ Lehar@ResponsibleAITrust.com

🌐 [@ResponsibleAITrust](#)

● [View the latest live version](#)



Join for updates on upcoming briefs, research insights, and governance resources.

 responsibleaitrust.substack.com

Powered by

