WHITE PAPER

tetrate

Your AI Is Already Out of Control: Why Governance Can't Wait

Table of contents

Introduction: Governance *Can* Be An Enabler

AI adoption is no longer optional—it's the fastest-growing priority in the enterprise. But governance is lagging behind, often taking the form of process-heavy, innovation-stalling frameworks that emphasize documentation over action. With AI's accelerating pace, organizational clockspeed becomes mission-critical.

The gap between ambition and execution is growing fast:

- $5M — Average GenAI project investment in 2024
- 10x — Potential miscalculation of GenAI costs
- 33% — GenAI proof-of-concepts expected to be abandoned by end of 2025

Executives are pouring millions into AI with unclear returns, no budget visibility, and few meaningful guardrails. Governance, if it exists at all, is often too slow to react—or too rigid to support real innovation.

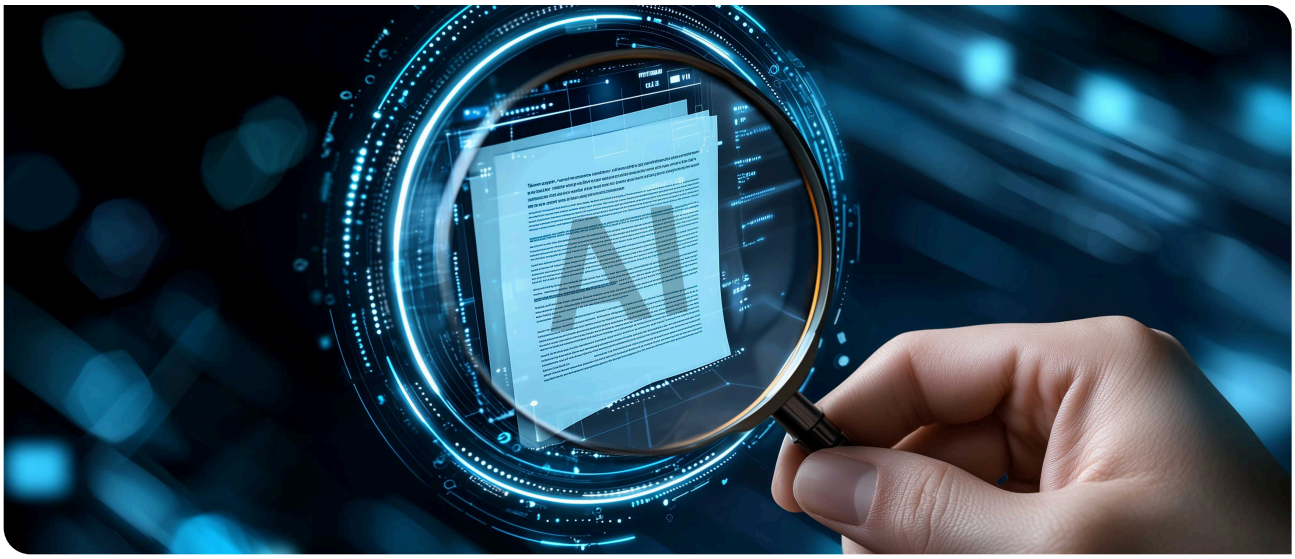
Meanwhile, shadow AI is spreading. Teams circumvent IT. LLM usage mushrooms without review. And when governance does appear, it often slows teams down—so they go around it.

> "Given that only 14% of organizations adopting AI have reached the highest maturity in AI governance, there is significant room for improvement across the enterprise landscape."
>
> *- Gartner, AI Maturity Matters:*
> *Mastering AI Governance and TRiSM Technology Drives Success, Avivah Litan , 6 June 2025*

We're past the point of hoping policies will work. To make AI safe and productive at scale, enterprises need **real-time control**, not just intention.

This whitepaper introduces a modern, action-oriented approach to AI governance—one that supports innovation without sacrificing visibility or safety.

Traditional Governance Isn't Enough—so What Comes Next



Most AI governance today falls into two camps. Each offers some benefit—but both fall short in meeting the urgency and scale of today's GenAI demands.

1. People Police: Governing Developers

This mode places the burden on developers to follow internal policies. Teams implement their own token caps, prompt restrictions, or output checks. But these guardrails often break down under pressure—especially when deadlines loom or incentives misalign. It's a flexible, developer-first approach—but one that's impossible to scale.

2. Passport Control: Governing Models

This approach tries to block risk at the front gate—limiting which models are accessible, wrapping APIs with heavy restrictions, or attempting to self-host commercial LLMs. While it offers centralized control, it slows down innovation, breeds resentment, and ultimately gets bypassed. The faster model innovation happens externally, the harder this gets to maintain.

Neither of these methods meets the moment for AI. Governance must be responsive and real-time in order to *enable* developers to use AI safely without slowing them down. Traditional methods rely on static policy, offline review, or manual compliance checks—none of which keep pace with the way GenAI is actually being used in modern enterprises.

3. Live Command & Control: Governing Traffic in Real Time

Introducing a third mode—what we call **Live Command & Control**—is fundamentally different. It observes AI traffic in real time, providing enforcement and visibility **in motion**. This is not a patchwork of developer rules or a perimeter-based blockade. It's a universal, dynamic, and resilient form of governance that:

- Applies policy mid-flight without interrupting workflows
- Offers complete visibility across orgs, clouds, and LLMs
- Enables showback, throttling, and auto-blocking of policy violations
- Treats GenAI usage like any other critical system workload

This isn't just a better control point—it's a new philosophy. Governance must be invisible until it matters, fast enough to keep up, and smart enough to adapt.

Why Live Command & Control Is the Missing Piece

Most AI governance efforts today focus on policy, not practice. Committees form, documents are written—but teams continue making real-time GenAI API calls with little oversight or alignment.

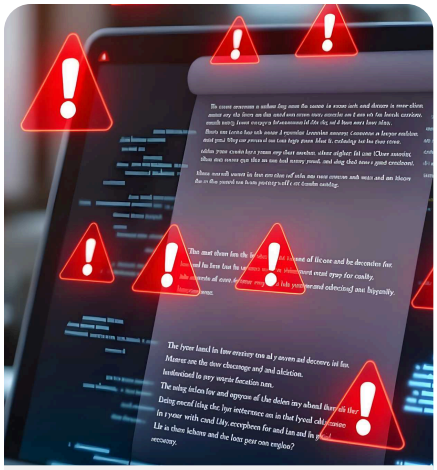
The problem isn't just a tooling gap. It's a strategic one.

Traditional governance methods—manual reviews, centralized approvals, security bottlenecks—slow teams down. GenAI adoption moves too fast, and business incentives are too strong. If controls add friction, they get bypassed.

Examples of failure:

A dev team routes prompts through a personal OpenAI key to avoid slow security reviews.

A business unit buys LLM access on a company card, creating unmonitored usage and cost.



A model wrapped in strict internal controls fails to meet product needs, so engineers revert to shadow tools.

> "Without implemented guardrails, AI models can rapidly generate compounding negative effects that spin out of control, overshadowing any positive performance and societal gains that AI enables."
>
> *— Gartner, Executive AI Governance Playbook, Svetlana Sicular, Nader Henein, et al., 8 January 2025*

To fix this, organizations need a control layer that operates in real time. One that doesn't depend on static reviews or developer goodwill. A system that can:

- Enforce policy without slowing teams down
- Provide visibility into AI usage across orgs and clouds
- Detect shadow AI before it causes damage
- Align guardrails with cost, risk, and business context

This is more than a technical fix—it's a governance philosophy. High-performing organizations don't just write policy—they operationalize it.

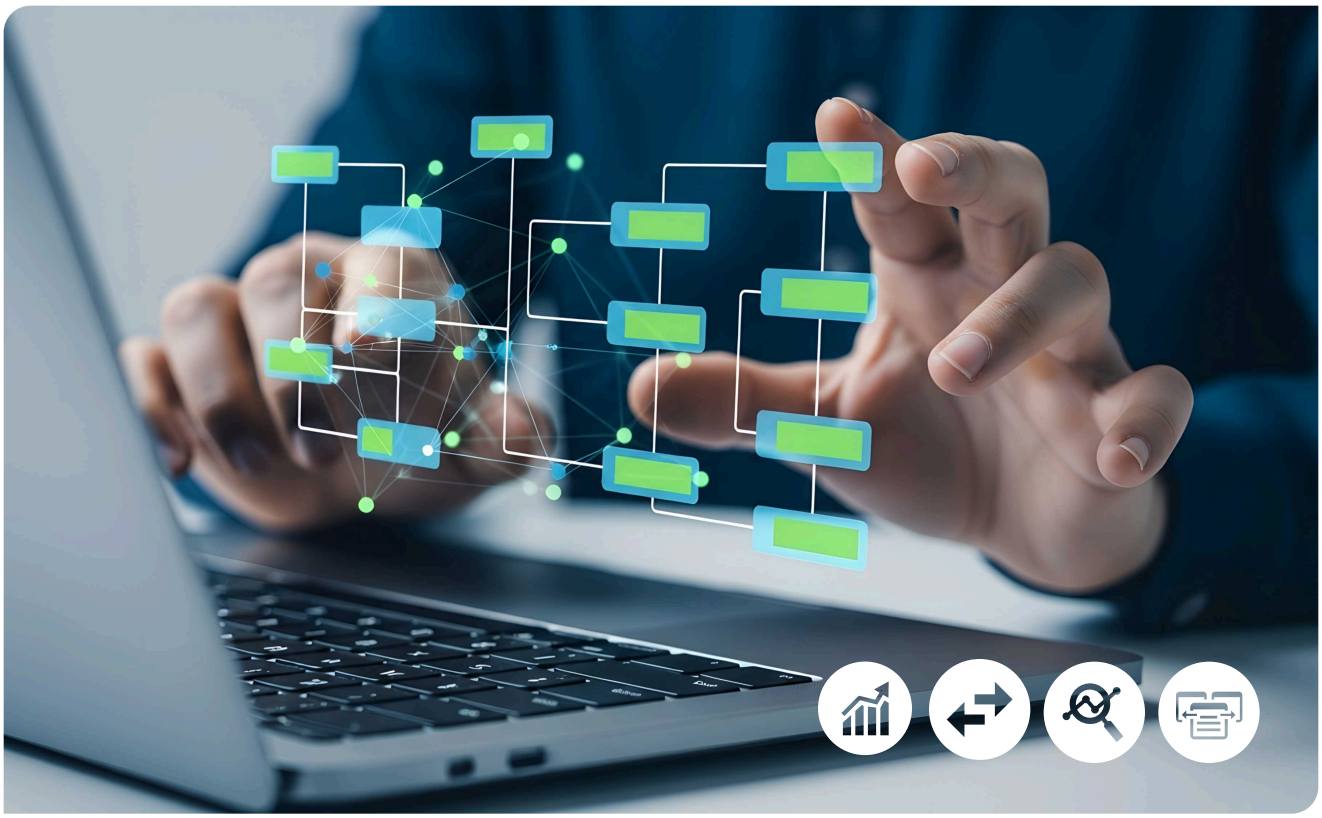
> "High-maturity companies overall achieve a 45% production rate for GenAI prototypes, compared to 31% for low-maturity organizations"
>
> — *Gartner, Gartner, AI Maturity Matters: Mastering AI Governance and TRiSM Technology Drives Success, Avivah Litan , 6 June 2025*

A strategic, real-time control layer turns governance from a blocker into a business enabler.

How Live Command & Control Enables Active Governance

Real-time governance turns governance from a passive checklist into an operational control system.

Rather than relying on static policy enforcement or disconnected team-level controls, real-time evaluation observes and manages AI usage in motion—across networks, clouds, and tools—without requiring developers to change code or teams to pause work.



This enables organizations to:

- Discover where GenAI is being used automatically
- Attribute usage to teams, apps, and business units
- Apply dynamic limits on spend, risk exposure, or usage volume
- Detect and block unsanctioned LLM traffic in real time
- Redirect or rewrite requests based on business policy
- Support showback or chargeback to increase accountability

With this capability, governance becomes embedded into the way work flows—unobtrusive but enforceable, flexible but firm.

Phases of an AI Transformation

This new way to govern doesn't happen all at once. Organizations must evolve their strategy in a way that enhances—not inhibits—business outcomes.

AI Transformation Journey

1. Ad Hoc Experimentation

- Reword: teams play with LLMs using company cards
- Tacitly allowed, no over-sight or standards

Road Blocks
- Cloud cost spikes
- Security/legal risks
- Quality inconsistency

Examples
- Code in GPT
- Rogue GPT use
- Code-gen blocked

Business Response
- AI banned
- Unstructured use

Governance: Help or Harm?
- No rules = slow, fearful
- Clear rules = safe speed

2. Siloed Acceleration

- BU leaders fund AI pilots; early cost-cutting wins and business cases emerge
- Revenue-generating POCs begin

Governance: Barrier or Enabler?
- Compliance slows
- Policing lags
- Controls delay

Business Response
- Policing stifles
- Crackdowns freeze

Examples
- Unvetted access
- Shadow AI stacks

Road Blocks
- Hidden costs
- Loss of control
- Shadow AI risks

3. Scaled Integration

- Central teams enforce LLM gateways and filters
- Proprietary models shared internally

Road Blocks
- Churn forces retools
- Costs spike, models shift models

Examples
- Spend untracked
- Filters bypassed

Business Response
- Secured access
- Ad-hoc AI slows

AI Governance Challenges
- Access curbs shadow AI
- Costly to run
- Governance drags
- Walled gardens strain

4. Enterprise Wide Real Time

- AI use requires approval, budget, and review
- Centralized teams monitor and control AI

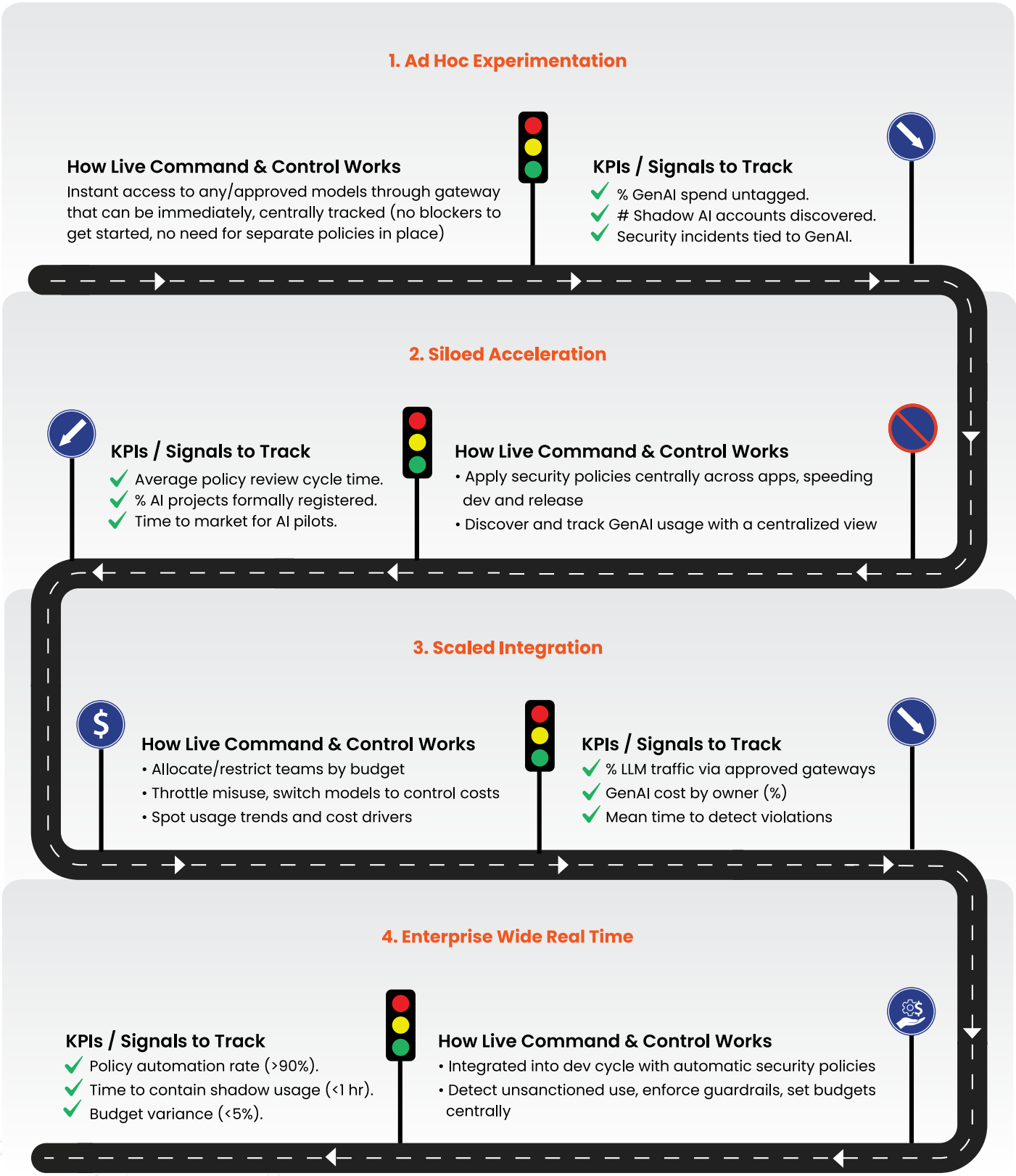
Governance vs. Agility in AI
- Controlled, safe tests
- Models lag
- Costs limit innovation

Business Response
- No free experiments
- Compliance delays

Examples
- Stops overspend fast
- Blocks shadow tools

Road Blocks
- Slow, less agile
- AI scale hinders leak tracking

The Value of Live Command & Control



Align Investments Across People, Process, and Technology

Moving from one phase to the next requires investment in governance maturity:

People

Create ownership structures and foster awareness.

Process

Embed governance into dev workflows and procurement paths.

Technology

Establish a real-time control plane that provides visibility and enforcement.

Key Steps to Advance

This isn't just a better control point—it's a new philosophy. Governance must be invisible until it matters, fast enough to keep up, and smart enough to adapt.

1. **Inventory Usage** – Identify and catalog where and how GenAI is used.
2. **Define & Publish Policy** – Align governance rules with actual business risk and opportunity
3. **Identify Ownership** – Connect usage to org structure, cost centers, and risk owners.
4. **Real-Time Showback** – Surface live dashboards and periodic reports that compare actual behavior against policy. Make overages, risky prompts, and runaway cost visible to the teams generating them—before central security steps in.
5. **Voluntary Remediation** – Give teams the data to fix issues themselves, preserving velocity while improving posture.
6. **Guardrail Enforcement** – Enforce policy dynamically without disrupting delivery.
7. **Iterate & Expand** – Continuously improve coverage and precision.

Consistent showback is what turns raw telemetry into behavior change. Transparency creates the peer pressure that shifts culture from "security says no" to "we own our footprint." This is how governance becomes a competitive advantage. Real-time control doesn't just solve for AI—it becomes a foundation for secure, adaptive operations across the enterprise.

Conclusion: The Time for Real Governance Is Now

AI is moving too fast for governance to be a policy exercise. It must be a platform capability.

The future of AI governance isn't about stopping adoption—it's about enabling it responsibly and at scale. Real-time control, powered by a strategic enforcement layer, is the path forward.

This isn't theory. It's the blueprint for making AI work—securely, sustainably, and successfully.

Don't Just Adopt AI—Govern It

Turn policy into platform capability

Contact Us

About Tetrate

Tetrate enables a safe and fast modernization journey for enterprises. Built atop Envoy and Istio, its flagship product, Tetrate Service Bridge, spans traditional and modern workloads so customers can get consistent baked-in observability, runtime security, and traffic management—for all their workloads, in any environment. In addition to the technology, Tetrate brings a world-class team that leads the open Envoy and Istio projects, providing best practices and playbooks that enterprises can use to modernize their people and processes.

Location: Tetrate, 691 S Milpitas Blvd, Suite 217, Milpitas, CA 95035, USA
www.tetrate.io | info@tetrate.io