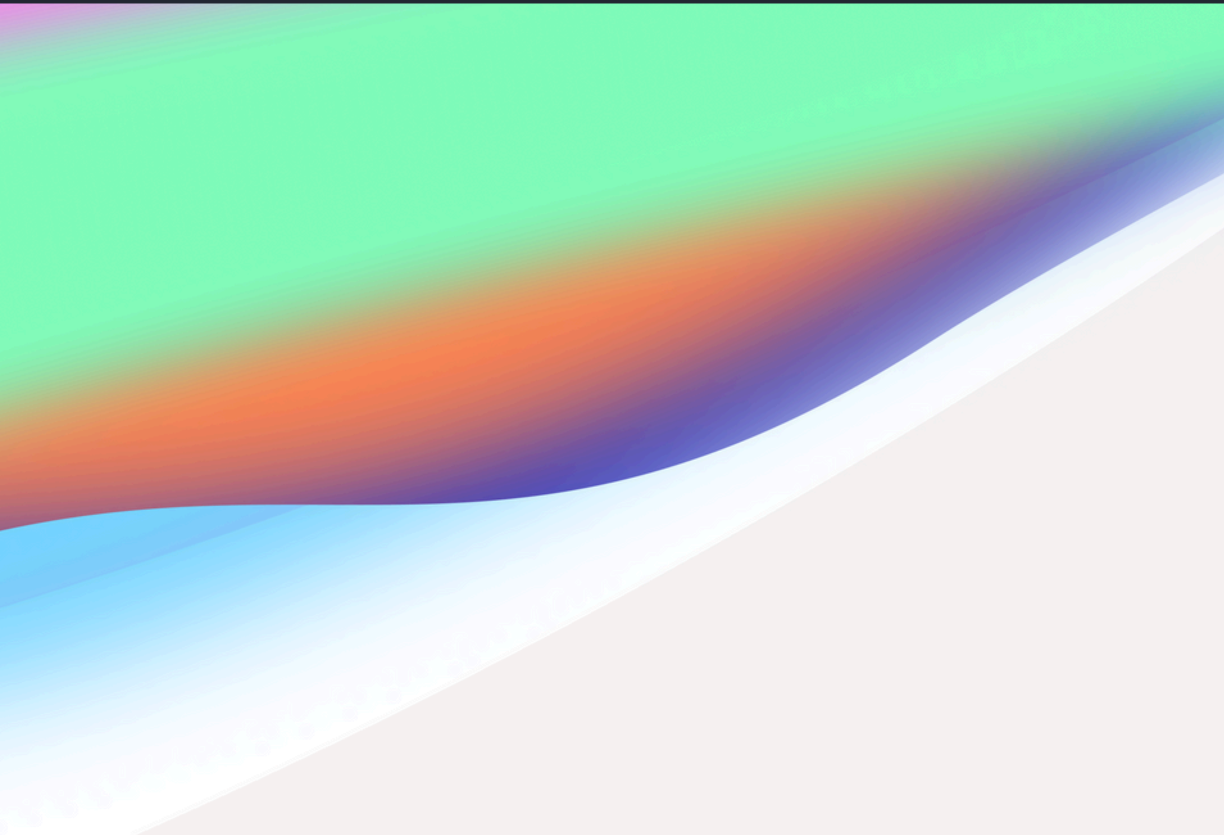
GenAI Policy Handbook 2025



Contents

2025 will be remembered as the year AI went from experimental to existential. Nearly every enterprise now engages with generative AI, whether they've approved it or not.

By 2025, more than 70% of enterprises will have implemented AI in at least one business unit, but fewer than 30% will have implemented formal security or governance policies tailored to GenAI.[1] Meanwhile, more than half of emerging cybersecurity startups now position AI as a critical threat surface.

The same systems that unlock business value are exposing organizations to unseen risks—hallucinations, data leakage, prompt injection, model abuse—and a shifting regulatory landscape.

This handbook was created to help leaders navigate that reality. It brings together the latest thinking on governance, security, and safe deployment of GenAI at scale — drawing from external experts and CalypsoAI's work with leading enterprises, governments, and builders.

Whether you're scaling adoption or just beginning your AI journey, we hope this serves as a trusted guide to securing innovation.

— Donnchadh **Casey, CEO, CalypsoAI**

Chapter 1

Introduction

Why GenAI Needs a Policy Now

Generative AI has shifted from experimental deployments to mission-critical infrastructure.

According to a recent survey of over 200 CISOs, 97% of CISOs say implementing AI tools is a strategic priority over the next 1-2 years—yet that same 97% also report concern over AI/ML-powered attacks.[2] The urgency to act is clear, but most enterprises still lack centralized controls, visibility and policy enforcement for GenAI.

At the same time, the complexity is compounding: new open-source models drop weekly, SaaS tools ship with AI capabilities baked in, and cross-functional teams—from developers to marketers—are experimenting independently. As model capabilities evolve, so do the attack surfaces.

GenAI isn't just software. It's a self-evolving, unpredictable system that requires new modes of control. This handbook is built to help you move from intention to implementation—bridging the gap between policy and security with practical, scalable guidance.

97%

of CISOs say implementing AI tools is a strategic priority over the next 1-2 years.

Chapter 2

Understanding the Risk Landscape

Today’s GenAI risks are often misunderstood, underestimated, or completely invisible until damage is done. Unlike traditional attack surfaces, **GenAI threats exploit the model’s behavior**, context, and interactivity—meaning even well-intentioned deployments can spiral into unintended outcomes.

A single prompt can lead to brand-damaging outputs, sensitive data exposure, or security failures that cascade through applications.

The Shifting Threat Surface

Dynamic, compound risk vectors include:

- **Input attacks** (prompt injection, data poisoning)
- **Output vulnerabilities** (PII leakage, hallucinated compliance violations)
- **Model misalignment** (jailbreaks, undesired behaviors)
- **Operational threats** (denial-of-wallet, inference layer exfiltration)

Unlike traditional software, generative AI systems interact continuously with users and content. That interaction surface is both your greatest innovation lever—and your biggest liability.

Agentic AI and the Rise of Autonomous Risk

As enterprises deploy agentic systems (AI that uses tools to complete tasks), risk becomes multi-dimensional:

- **Unbounded behavior:** Autonomous agents making financial, legal, or HR decisions
- **Chained logic errors:** AI acting on outputs from other AIs
- **Model Context Protocols (MCPs):** New integration pathways for attack entry

With Agentic AI set to arrive in the enterprise faster than most security programs can adapt, security leaders must now be prepared to secure machine identities, monitor emergent behaviors, and red-team agentic activity to prevent dangerous decision loops.[3]

Chapter 3

Controls & Security Layers for GenAI

Why Controls Must Center on Inference

The apex of AI innovation—foundation model training—is controlled by a few hyperscalers who wield immense computational resources and can navigate the complexities of massive datasets.

The vast majority of enterprises operate at the **inference layer**, where real work (and real risk) happens. Inference is where:

- Users interact with models
- Data is exposed or ingested
- Outputs affect business outcomes

Securing AI at the inference layer is critical and requires a holistic approach that encompasses three core areas:

1

Defense

Establish resilient defenses by applying input sanitization, output filtering, and runtime monitoring tailored to each use case. Controls must account for both direct threats and cross-pipeline vulnerabilities, adapting in real-time to evolving risk.

2

Offense

Use proactive red teaming to uncover weaknesses in both models and their applications. Through adversarial testing and agentic attacks, organizations can identify, score, and prepare for emerging attack paths before they're exploited.

3

Governance, Regulation & Compliance:

Map evolving regulatory frameworks to practical controls. Rather than relying solely on "paper policies," translate compliance requirements into enforceable guardrails—ensuring policies are lived, not just written.

This approach minimizes the attack surface and allows for robust security measures to be implemented around the model and application, regardless of the model's internal complexities.

Core Controls to Prioritize

To achieve a holistic strategy, layered protections, dynamic enforcement, and continuous adaptation is required. The following categories represent foundational building blocks for securing GenAI at the point of interaction — the inference layer — where users, data, and models converge. These controls should be customized based on organizational risk appetite, use case sensitivity, and model complexity.

Input Controls

- Prompt sanitization
- Forbidden term filters (e.g., internal project names, employee data)

Output Controls

- Toxicity filtering
- PII redaction
- Response scoring (alignment confidence)

User Role Management

- Role-based access to models
- Tiered use-case approval (e.g., low-risk marketing chatbot vs. finance agent)

Runtime Observability

- Logging every interaction
- Anomaly detection (e.g., excessive prompt length, unusual access patterns)

Red-Teaming for AI

- Signature attack libraries (e.g., jailbreak prompts)
- Agentic Warfare (privilege escalation, denial-of-wallet)
- Risk scoring for models & applications
- AI Agent Risk Management (agent-specific runtime security controls, identity governance for nonhuman identities, and policy enforcement frameworks that extend to multiagent systems).[1]

Shadow AI Detection

- Discover unauthorized AI tool usage
- Monitor for data egress via GenAI interfaces

Customizable Control Layers

- Implement unified security platforms that support customizable controls — the ability to adjust policy enforcement to specific needs (i.e., user role, model type, use case, and sensitivity).
- Supports both conservative and innovation-oriented teams within the same organization.

> These controls form the basis of a modern AI policy framework —one that focuses not just on documentation, but on living, adaptable, and enforceable practices.

Chapter 4

Building a GenAI Policy Framework

Make It Real, Not Rigid

Policies are only useful if they are:

- **Context-aware** (tailored to use case, data sensitivity, model risk)
- **Operationally enforceable** (can be embedded in workflows and tooling)
- **Adaptable** (updated as models or integrations evolve)

Key Policy Domains to Cover

Acceptable Use

- Prohibited GenAI uses (e.g., legal advice, financial forecasting without oversight)
- Authorized model providers and plugins

Prompt Guidelines

- Redline prompt inputs (e.g., “Do not include PII,” “Do not recreate copyrighted content”)
- Templates for safe prompting

Data Classification & Protection

- Data tiering (public, internal, sensitive, regulated)
- Controls for regulated inputs (e.g., healthcare, finance, personal data)

Model Access Policies

- Permissions by role and department
- Approval processes for new models or integrations

GenAI governance is about building adaptive, enforceable policies that reflect how it’s being used, who’s using it, and what’s at risk.

Testing & Evaluation

- Red-team thresholds before production use
- Scoring systems for model security and alignment

Logging, Retention & Incident Response

- Required logging of model interactions
- Escalation paths for prompt or output incidents
- AI-specific incident response mapped to SOC procedures

Third-Party & SaaS AI Governance

- Procurement policies for AI-enabled software
- Vendor security assessments and risk tiers

Employee Awareness & Training

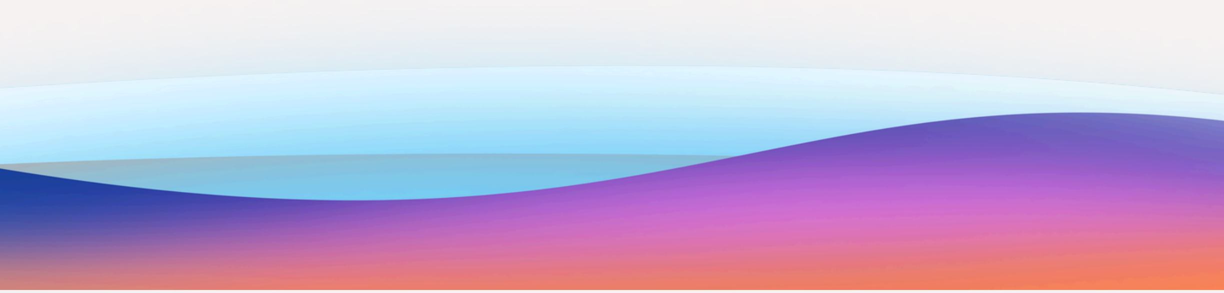
- Role-specific AI usage education
- Safe prompting workshops and regular refreshers

Chapter 5

Governing AI

Navigating the evolving regulatory landscape is a critical aspect of AI security. Jurisdictions and industry sectors are rolling out GenAI-specific regulations—ranging from the EU AI Act to sector-specific guidance in finance, healthcare, and government.

Compliance cannot exist in isolation.



Governance, Regulation & Compliance

An effective Governance, Regulation and Compliance (GRC) approach must bridge the gap between static documentation and active enforcement. This includes:

- Mapping legal and regulatory obligations to technical controls
- Ensuring auditability of AI use and decisions
- Implementing retention policies that meet sector standards
- Demonstrating policy compliance during internal or external reviews

Crucially, compliance should not exist in isolation. Your GRC posture should be integrated with policy frameworks, runtime controls, and governance structures to ensure defensibility and agility.

Governance as Continuous Oversight

Traditional governance is retrospective. GenAI requires continuous oversight across lifecycle stages:

- Model selection & usage
- Prompt construction & ingestion
- Output review & escalation
- Monitoring, logging, and tuning

Governance Structures to Consider

- **AI Steering Committees:** Blend legal, security, innovation, data, and business roles
- **Risk Assessment Panels:** Score and review use cases before deployment
- **Federated Champions:** Business-unit leaders who help scale policy adherence

Financial Governance

Gartner predicts that by 2027, 60% of enterprises will adopt Financial Operations practices to manage AI cost unpredictability.[1] Therefore, AI-specific FinOps disciplines should be embedded into your governance strategy to track, forecast, and optimize spending across model usage, infrastructure, and orchestration workflows. Failure to implement robust financial controls is now considered a leading risk to GenAI initiatives.

60%

60% of enterprises will adopt FinOps practices to manage AI cost unpredictability

Governance should not stifle innovation—it should provide clear, trusted lanes for secure adoption.

Tool: Example RACI Chart

Task	Responsible	Accountable	Consulted	Informed
AI Use Case Approval	Head of AI	CDO/CPO	Legal, CISO	Line Managers
Prompt Security Review	Security	CISO	DevOps	All Developers
Incident Investigation	SOC	CISO	Legal, Privacy	Executive Team
Policy Update & Distribution	CDO	CIO	Legal, HR	All Staff

Chapter 6

Maturity Model

& Rollout Roadmap

Where Are You Now?

Assess your current state across three tiers:

Maturity Level	Description	Traits
Reactive	Uncoordinated usage, no visibility	Shadow AI, ad hoc rules, audit risk
Proactive	Policy & controls in place	Runtime enforcement, red-teaming, AI use registry
Optimized	Fully integrated, adaptive governance	Dynamic risk scoring, inference-layer observability

12-Month Roadmap Example

Quarter	Focus Area	Key Activities
Q1	Inventory & Baseline	Model audit, stakeholder alignment
Q2	Core Policy & Controls	Input/output filters, red-team, role-based use
Q3	Governance Setup	Launch steering group, set review cadence
Q4	Scaling Secure Adoption	Expand to new use cases, federated champion model

Chapter 7

Recommendations

& Resources

Top Questions to Ask Now

1. What models are being used across our organization?
2. Who can access GenAI tools and for what use cases?
3. Do we log, audit, and analyze AI interactions?
4. Have our models been red-teamed for security and misuse?
5. Do we have acceptable use policies tied to enforcement?
6. Are we blocking unsafe prompts or responses at runtime?
7. Do we have visibility into shadow AI?
8. Are roles and permissions tailored by risk level?
9. Can we demonstrate responsible usage to regulators?
10. Who owns AI governance—and how often is it updated?

Glossary

Inference Layer: *The point where AI models are queried and produce responses*

Shadow AI: *Unauthorized or unmanaged AI use within an organization*

Prompt Injection: *Manipulating a model’s behavior through carefully crafted inputs*

Agentic AI: *AI agents capable of taking autonomous actions with tools*

Agentic Warfare: *Using empowered AI agents to automate red-teaming of AI systems for weaknesses and flaws*

Additional Resources

Gartner AI TRiSM Framework (2025)

ISO/IEC 42001 (AI Management Standard)

CalypsoAI Security Leaderboard

NIST AI Risk Management Framework

EU AI Act Overview & Timeline

Lightspeed Venture Partners & Wakefield Research Cyber60 Report & CISO Survey 2024-2025

Sources

1. *Gartner Predicts 2025: AI's Impact on the Future of Enterprise Technology*

2. *Lightspeed Venture Partners & Wakefield Research Cyber60 Report & CISO Survey 2024-2025*

3. *Forrester Top Recommendations For Your Security Program, 2025*