Protecting AI, Driving Growth: The Business Case for CalypsoAI

Contents

Executive Summary

“*There are only two types of companies: Those that have been hacked and those that will be hacked.*”
Robert S. Mueller, III, Former FBI Director

This quote remains relevant today as organizations increasingly rely on AI to drive business value. While AI offers significant opportunities, it also introduces new and less visible risks that traditional security programs aren’t designed to catch.

The consequences of misuse or compromise grow as AI becomes more embedded in decision-making and operations. A piecemeal approach to AI security is no longer sufficient. A comprehensive, organization-wide strategy is needed to manage these emerging risks, protecting not just systems, but reputation, trust, and long-term resilience.

The Problem

Today's security tools were built for traditional IT systems (deterministic software with fixed logic), not adaptive, autonomous AI systems that reason, generate, and act unpredictably. As enterprises integrate GenAI and agents into business workflows, they are facing entirely new challenges, including:

AI-Specific Threats

Prompt injection, jailbreaks, data leakage, and model manipulation are not addressed by traditional cybersecurity tools.

Compliance Complexity

Regulations like the EU AI Act or ISO/IEC 42001 demand explainability, traceability, and bias detection. Enterprises need solutions that evolve with global policies.

Inference Blind Spots

Most organizations lack visibility and enforcement at the point where AI decisions are made. This point is called inference, and it is where risk manifests.

Operational Pressure

AI teams are under pressure to deploy fast. Security teams are under pressure to slow them down. This creates gridlock.

This new reality is creating enormous pressure on security teams. They're being asked to protect AI systems they often didn't build, don't fully understand, and can't always trace. This is compounded by an explosion in generative and agentic AI use cases across the enterprise, which while driving innovation, also introduces a new class of security risks.

A Snapshot of the Evolving Risk Landscape

Risks of Selecting the Wrong Model

Most organizations choose models based on performance and cost, but fail to evaluate the security posture of those models, leaving them open to:

Prompt Injection & Jailbreak Vulnerabilities

Many models can be manipulated into producing harmful or unauthorized outputs through cleverly crafted prompts, exposing sensitive data, legal liabilities, or reputational damage if outputs are inappropriate, biased, or unsafe.

Data Leakage

Some models are built on sensitive training data or confidential inputs, posing privacy and IP exposure risks.

Hidden Backdoors

Undocumented or adversarial behaviors—intentionally or unintentionally embedded during training—can be triggered post-deployment.

Poor Access Controls

Insufficient authentication and authorization mechanisms can lead to unauthorized use and misuse of models or applications.

Undocumented Capabilities

Some models can produce malicious code or bypass filters in ways that only surface under real-world conditions.

Product Deployment Risks

Once deployed, models face a host of runtime threats that standard application security tools are not equipped to handle:

97%

of participants say they have encountered breaches or security issues related to the use of GenAI in the past year alone.

Model Extraction

Attackers can reverse-engineer model behavior by querying it systematically, stealing proprietary AI capabilities.

Evasive Jailbreaks & Prompt Injections

Sophisticated prompt engineering can bypass model guardrails and content filters.

Indirect Prompt Injection

Malicious inputs embedded in legitimate data fields (like user-submitted tickets) can trigger unintended model behavior.

Privacy Violations

Outputs may unintentionally reveal patterns tied to individuals, violating GDPR, CCPA, and other privacy laws.

Operational Disruption

Attacks that degrade model performance or availability (e.g., through resource exhaustion or misalignment).

Compliance Failures

Inappropriate outputs may breach regulatory standards, especially in financial, healthcare, or legal contexts.

Adversarial Inputs & Social Engineering

AI-enhanced attacks can manipulate both the models and their human users.

These risks are already impacting enterprises. A survey conducted by the Capgemini Research Institute revealed that 97% of participants say they have encountered breaches or security issues related to the use of GenAI in the past year alone.

The Solution

The CalypsoAI Inference Platform addresses these threats holistically through agentic red-teaming, real-time defensive controls, and enterprise-wide observability, giving organizations the layered security gateway required to deploy GenAI securely and confidently across the entire AI lifecycle.

CalypsoAI's unified AI security platform is:

Model agnostic, working seamlessly with any LLM, public or private, across diverse applications.

Policy-driven to enforce security policies and access controls that prevent unauthorized model access.

SIEM and SOAR compatible for seamless enterprise integration.

Scalable with ultra low-latency across large deployments to maintain user experience.

Trusted and proven reliability by Fortune 500 and national security customers to secure high-volume, sensitive deployments

Expert-backed with over 70 years of combined AI security expertise, research and around-the-clock threat monitoring.

The products that sit on the CalypsoAI Inference Platform create a purpose-built, enterprise-ready gateway that protects generative and agentic AI systems across all models, vendors, and environments.

CalypsoAI Inference Security Solutions

Inference Red-Team

Capabilities

01

Agentic Warfare™

leverages CalypsoAI's attack agents to dynamically red-team AI systems based on custom intents that align to specific use cases and risks. Waging Agentic Warfare™ on systems in this way enables organizations to detect vulnerabilities that only emerge during complex, multi-turn interactions.

02

Continuously updated, extensive Signature Attacks

are out-of-the-box, curated and tested single-turn prompts, updated monthly with 10,000+ new prompt packs that target high-severity intents derived from real-world insights. Signature Attacks enable enterprises to test their entire AI stack upon deployment for rapid vulnerability detection.

03

Operational Attacks

force errors, resource consumption, latency issues, crashes, and denial-of-service or denial-of-wallet application attacks formulated specifically for AI. This level of operational red-teaming enables security teams to show that AI systems undergo the same level of rigorous performance testing as traditional applications, giving you visibility into potential system failures and the confidence to address operational risks before impacting business continuity.

All Inference Red-Team attacks can be automated and generate a detailed report that includes malicious prompts, model responses, security scoring, and severity classifications to prioritize remediation.

Secure Model Selection

Inference Red-Team reports include two security scores that help you find the safest model and stress test your AI system before deployment.

01 The CalypsoAI Security Index (CASI) score gives a top-level view of a model's general security posture and how resilient it is to a broad spectrum of vulnerabilities. CASI doesn't just count attack success rates; it weighs the severity of successful breaches, the complexity of attack paths, and the defensive breaking point, where a model's guardrails start to fail.

02 The Agentic Warfare Resistance (AWR) score takes it a step further by assessing how a model can compromise an entire AI system. With a specific intent, our trained attack agents plan, iterate and adapt in order to:

- Extract user-provided system prompts
- Break the model's alignment based on those system prompts
- Extract sensitive PII when the model is integrated into a RAG system

Together, CASI and AWR provide the industry's most comprehensive, proactive view of AI model risk.

Use Cases

Unvetted AI Model Selection

Teams adopt AI models without fully assessing their security risks and suitability for enterprise use.

Inference Red-Team

- Evaluates model security by identifying vulnerabilities before deployment.
- Ensures suitability for use case by assessing risks based on organizational needs.
- Reduces risk exposure to prevent the adoption of unsafe or unreliable models.

Insecure AI Development & Deployment

AI-driven applications are built without security testing, increasing vulnerabilities throughout the SDLC.

Inference Red-Team

- Tests AI applications for vulnerabilities by identifying weaknesses before production.
- Strengthens AI security posture by reducing risks in AI development workflows.
- Ensures compliance from the start by aligning AI security with regulatory standards.

Rapidly Evolving AI Threats & Attacks

New AI-specific vulnerabilities emerge constantly, requiring continuous testing to stay secure.

Inference Red-Team

- Continuously tests AI and agentic systems by integrating security into CI/CD pipelines.
- Detects emerging threats by informing defensive controls for the strongest security posture.
- Enhances AI resilience by ensuring models remain secure over time.

Outcomes

Maximum Efficiency with Minimal Resources

Automated AI security testing eliminates the need for specialized teams, freeing resources for priority initiatives.

Accelerated Time-to-Value

With same-day setup, vulnerabilities can be discovered in as little as minutes, delivering actionable insights immediately.

Proactive Detection

Stay ahead of threats by uncovering vulnerabilities early with continuously evolving, agentic attack scenarios.

Efficient Collaboration and Reporting

Share clear, actionable findings that streamline handoffs across teams and prioritize high-impact issues for faster remediation.

Automated Assessments

Automated, scheduled assessments ensure AI defenses stay ahead of evolving threats on a continuous basis.

Inference Defend

Capabilities

01
Out-of-the-box scanners

ship with pre-configured scanner packages for PII, prompt injection, jailbreaks, and adversarial content. These packages are constantly updated and provide instant security coverage from deployment, ensuring fast time-to-value with enterprise-grade accuracy.

02
Custom scanners

can be calibrated to company-specific terminology, unique risk profiles, or time-sensitive use cases. These are easily managed through CalypsoAI's secure Playground interface and can be published and grouped into reusable packages for ease of use.

03
Protect not obstruct

so teams can dial protections up or down based on context, enabling strict moderation where needed and lighter-touch oversight where flexibility is essential. Block, audit, or customize scanner responses to align with business and regulatory needs.

Inference Defend blocks 97% of harmful prompts with 95% decision accuracy, correctly identifying 92% of blocked prompts as threats. Backed by a 5x improvement in scanner latency, Inference Defend's security controls deliver an elite level of protection without compromise.

Inference Observe

Capabilities

Coupled with Inference Defend, Inference Observe eliminates blind spots, delivering enterprise-wide oversight to track AI usage, detect threats, and enforce compliance without disrupting workflows through:

01
Unified monitoring and reporting

which enables real-time AI security insights with detailed logging and audit trails, giving security teams the visibility needed to track, investigate, and mitigate risks.

02
Global dashboards

that provide a centralized view of AI usage and security events to enable compliance across the entire enterprise.

03
Automated policy flagging

ensures AI applications that do not adhere to corporate governance, regulatory requirements, and internal security frameworks are identified.

Use Cases

Sensitive Data Exposure

AI interactions can unintentionally expose personal, financial, or proprietary data, leading to breaches and compliance risks.

Inference Defend

- Prevents data leaks by scanning for and blocking sensitive information before it reaches AI systems.
- Ensures regulatory compliance by protecting against violations of data protection laws.
- Safeguards proprietary information by securing trade secrets, legal documents, and customer data.

AI Manipulation & Security Threats

Adversaries exploit AI through prompt injections and jailbreaks, bypassing safeguards to extract confidential data or generate harmful outputs.

Inference Defend

- Blocks prompt injection and jailbreak attacks by detecting and stopping malicious AI manipulations.
- Secures AI interactions by ensuring AI responds safely and ethically.
- Prevents unauthorized access by monitoring and controlling AI usage within organizations.

Business-Specific Risk Management

AI applications often have different thresholds of risk tolerance depending on function, audience, and business objective. Security teams need tailored controls that reflect these nuances.

Inference Defend

- Enables custom scanners and flexible policy controls to align protection with distinct use case needs.
- Supports precise security configurations that reflect business risk appetite, compliance frameworks, and operational context.
- Ensures AI applications remain protected without overblocking or disrupting intended functionality or innovation initiatives.

Lack of AI Oversight & Visibility

Organizations struggle to track and control AI usage, increasing risks of security gaps, inefficiencies, and non-compliance.

Inference Defend

- Monitors AI usage in real-time by providing visibility into AI interactions across teams.
- Enhances security and risk management by identifying anomalies and potential threats.
- Optimizes AI governance by enabling policy enforcement for AI applications.

Outcomes

Real-Time AI Security

Stops malicious inputs, harmful outputs, and unauthorized interactions before they cause damage.

Centralized Enforcement

AI security, policy enforcement, and compliance from a unified platform.

Seamless AI Performance

Security that works in the background, ensuring AI remains fast, reliable, and uninterrupted.

Adaptive & Future-Proofed

Continuously updated security policies and defenses to stay ahead of evolving threats.

Comprehensive Threat Detection & Response

Monitors, analyzes, and responds to AI security threats in real-time, preventing breaches before they escalate.

Return on Investment for CalypsoAI

An Analysis of Cost Impact Across Three Core Vectors

AI Security Impact

Labor Efficiency Gains

Automating manual compliance and monitoring tasks can lead to significant savings in labor costs (i.e., reducing the need for manual data entry and report generation).

Incident Cost Reduction

Breaches resolved within 200 days cost $3.93M; those beyond cost $4.95M. CalypsoAI helps close this gap, saving $1.02M annually by accelerating AI threat detection and containment.

3. Red-Team Intelligence

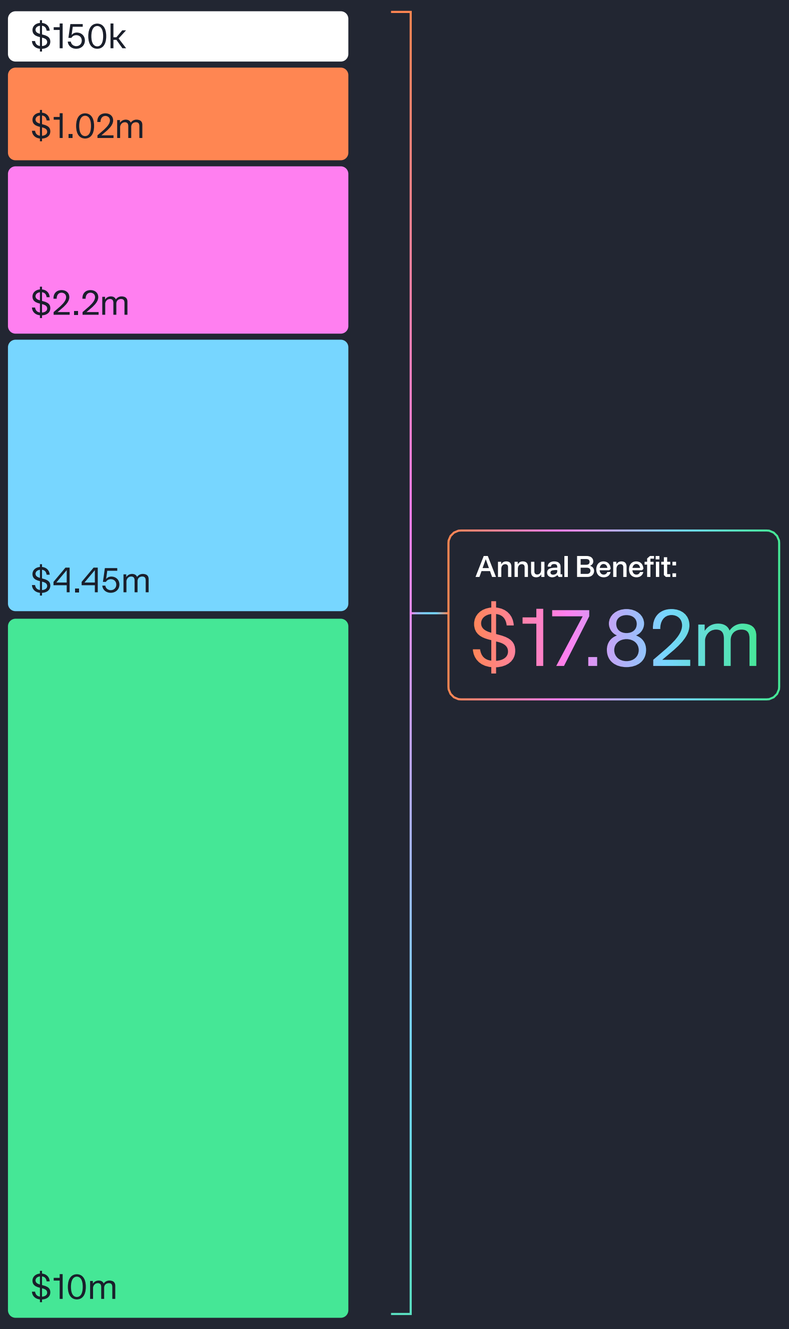
Proactively identifying and addressing vulnerabilities through advanced red-teaming and observability saves organizations significant cost compared to the cost (and time) of more manual red-teaming options.

Breach Risk Reduction

The global average cost of a data breach in 2024 was $4.45 million. Implementing robust AI security measures can prevent such breaches, directly saving this amount annually.

Regulatory Fine Avoidance

Non-compliance with the EU AI Act or GDPR can lead to fines up to €35M or 7% of global revenue. Avoiding these and legal costs can conservatively save $10M+ annually.



Brand Impact

Avoid Negative Press & Regulatory Scrutiny

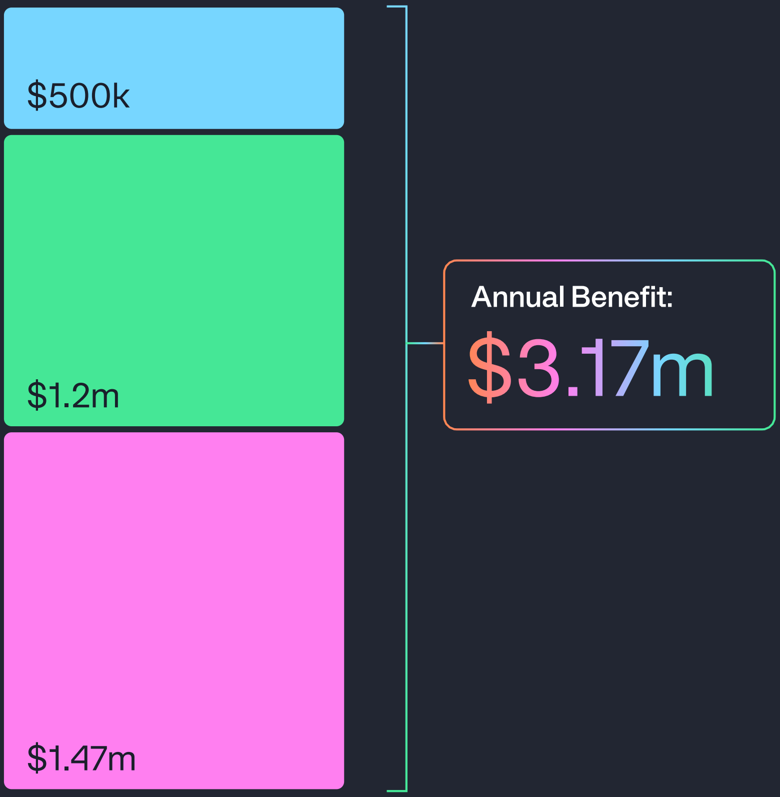
AI failures can trigger PR crises, market losses, and investigations. Proactive security helps prevent these events and their reputational and financial consequences.

Responsible AI Differentiation

AI adopters increase revenue 1.5x faster than competitors. Secure AI deployments can drive a conservative $1.2M annual revenue increase through trusted innovation.

Preserve Trust & Reputation

The average cost of reputation damage or loss of revenue due to a data breach in 2024 was $1.47 million. Maintaining strong AI security helps preserve brand trust, avoiding this potential loss annually.



Innovation Enablement

Accelerated AI Adoption

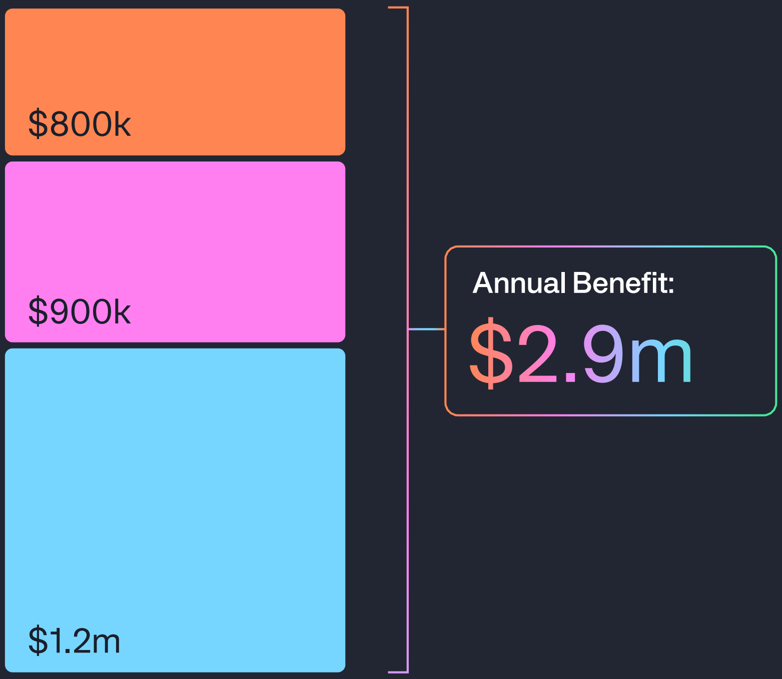
Secure AI reduces deployment delays. 67% of leaders report 25%+ revenue growth from AI. Faster launches amplify this impact.

Increased Efficiency & Capacity

Secure automation streamlines workflows and reduces errors, saving up to $900K annually while improving accuracy and productivity.

Competitive Differentiation

Organizations leading in secure AI adoption grow revenue 1.5x faster than peers. This trust-driven advantage delivers a conservative $1.2M in annual gains through scalable, responsible innovation.



Regulatory Drivers of ROI

Regulations like the EU AI Act raise the stakes:

- Fines up to $30m or 6% of global annual turnover.
- Requirements for continuous monitoring, explainability, and risk management.

CalypsoAI enables:

- **Automated compliance tasks** (saves up to $500K/year).
- **Avoidance of fines** ($10M+ in some scenarios).
- **Faster time to value by securing projects at launch.**

Build Your Own vs. CalypsoAI

	BYO	CalypsoAI
Time to Deploy	6–12 months	< 1 month
Total Cost of Ownership	High & unpredictable	Fixed & efficient
Compliance Readiness	Limited	Built-in
Security Breadth	Guardrails only	Full inference coverage
Innovation Impact	Slow deployment	Accelerated delivery

The Takeaway

With Calypso AI You Get

Security that goes first so your **innovation** can go **further.**

$17.82m

Real **ROI** with **$17.82m** in annual directional **benefits.**

Unmatched AI security with CalypsoAI's agentic red-teaming and **real-time defense** in a unified platform.