WHITE PAPER

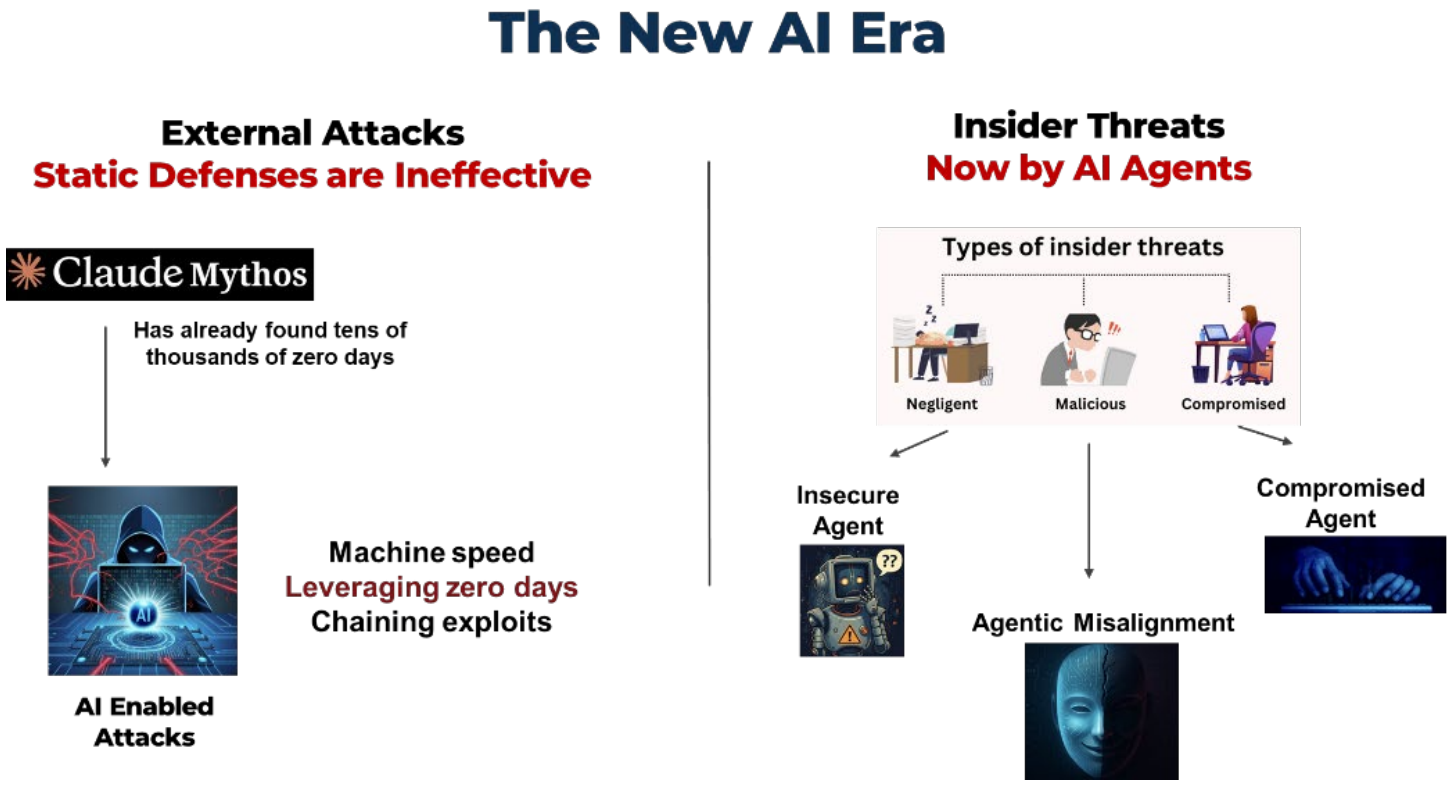
ACALVIO
AI-POWERED DECEPTION

Breaking AI-Attack Automation in the Generative AI Era

Acalvio delivers comprehensive, enterprise-wide preemptive cyber-defense. Powered by our next-generation 360 Deception platform, our deployments seamlessly span IT and OT environments, on-premises datacenters, and multiple cloud architectures. This deception covers both internal networks and external-facing perimeters, providing a proactive shield across the entire enterprise.

The Three Types of Generative AI Threats

As organizations and attackers adopt Generative AI, the attack surface has fundamentally shifted, and breakout times have collapsed to minutes and seconds. Frontier models like Claude Mythos are able to discover zero-days and string exploits to compromise enterprises at unprecedented machine speed. Because AI attackers are now operating faster than traditional patching, anomaly detection, and remediation cycles can handle, reactive defense is structurally late. Preemptive deception is becoming the only viable defense.



The Three Types of Generative AI Threats (continued)

We broadly categorize the modern GenAI threat landscape into three types:

1. **AI-Driven Attacks from the Outside Adversaries** are weaponizing AI to launch highly sophisticated, automated campaigns. These external threats leverage generative models to dramatically accelerate reconnaissance, deploy payloads, and autonomously string together exploits before defenders can react.
2. **Insider Threats by AI Agents** As autonomous AI agents (such as those built on the OpenClaw framework) are integrated into enterprise workflows, they introduce a new paradigm of insider risk. This includes agentic misalignment, the malicious use of AI agents by authorized personnel, and systemic misconfigurations that grant agents excessive permissions.
3. **Attacks Against AI Infrastructure** The underlying infrastructure powering GenAI has become a primary target. Attackers directly target the models, retrieval pipelines, and connective tissues of enterprise AI applications to steal proprietary data or hijack system outputs.

Acalvio's AI Threat Defense Strategy

Traditional security controls struggle to keep pace with autonomous adversaries. To combat these specialized threats, Acalvio utilizes **360 Deception** to destabilize attacker automation. By corrupting what AI can confidently trust, we force exposure before privilege escalation or domain compromise.

We map specific, advanced capabilities to each attack category:

For External AI-Driven Attacks: We deploy 360 Deception to create a high-uncertainty environment that disrupts the stable ground truth automated attacks depend on. In addition, our external-facing deception (ShadowPlex TTI) provides the first layer of high-fidelity intelligence, forcing exposure and diversion while gathering early threat intelligence on intrusion attempts.

For Insider Threats by AI Agents: We utilize a dual approach: injecting specialized Traps inside AI agents to catch malicious behavior at the source, backed by our 360 Deception coverage to detect silent lateral movement across the network.

Gartner
Dec 2025

For Attacks on AI Infrastructure: We deploy specialized AI Infra Decoys that mimic the actual generative AI technology stack, acting as an early warning layer when attackers attempt to probe AI pipelines.

Deep Dive: Core Deception Capabilities

360 Deception:

Disrupting the "Stable Ground Truth" Our 360 Deception engine breaks AI attack automation by creating multiple layers of uncertainty, forcing the adversary to verify more, trust less, and move slowly. This is achieved through:

- **Dynamism & Evolving HoneyPaths:** We render AI-driven environment mapping obsolete through continuous shifting. HoneyPaths are deceptive routes that guide attackers toward controlled environments. Because these paths evolve over time, they ensure that the map an adversary generated yesterday is entirely unreliable today.
- **RLF (Real Looking Fake) & Cloaking:** Traditional deception relies on fake assets that look real. 360 Deception expands this by simultaneously cloaking production assets to appear deceptive, and introducing intentionally suspicious artifacts that cannot be safely ignored. When real systems look like decoys to an AI, the attacker's machine-speed advantage is canceled.

Traps Inside AI Agents (Honeyskills)

To combat rogue, misaligned, or exploited enterprise AI agents, Acalvio is pioneering agent-centric deception. We have developed specialized **"Honeyskills"—**decoy tools, API endpoints, or functions injected into an agent's workflow that trigger an immediate high-fidelity alert when a hijacked or misaligned agent attempts to execute them. Initially developed to target the skills system (e.g., SKILL.md directories) of the popular OpenClaw agent framework, we are actively extending this capability to trap anomalous behaviors in major models like Anthropic's Claude Mythos and other leading AI agents.

AI Infrastructure Decoys

Protecting the connective tissue of enterprise AI requires specialized decoys that mimic the actual AI stack. Acalvio deploys high-fidelity infrastructure decoys including:

- **Decoy Chatbots:** Simulated conversational interfaces designed to detect prompt injection and system enumeration.
- **Decoy RAG Pipelines:** Fake Retrieval-Augmented Generation databases designed to trap attackers seeking to exfiltrate proprietary corporate knowledge bases.
- **Decoy MCPs:** Simulated Model Context Protocols to flag unauthorized API requests and trap autonomous agents attempting lateral movement across enterprise systems.

The Bottom Line:

In an environment where attackers operate faster than human defenders can react, trying to outpace them is a losing battle. The only viable defense is to control the reality and corrupt the AI's confidence, forcing malicious automation to verify more, trust less, and expose itself before it can escalate.

Acalvio is an AI-powered preemptive cybersecurity company focused on countering AI-driven identity and infrastructure intrusion. Its 360 Deception platform combines Dynamic Deception, evolving HoneyPaths, and cloaking of production assets within deception fabric to disrupt automated reconnaissance, credential abuse, and lateral movement across identity systems, endpoints, cloud, network, and cyber-physical environments. By altering what attackers can perceive and trust, Acalvio shifts detection from post-compromise analysis to pre-impact exposure, enabling organizations to detect, delay, disrupt, and deny malicious activity at machine speed. The company serves enterprise and government organizations determined to break automated intrusion at its source.
www.acalvio.com/