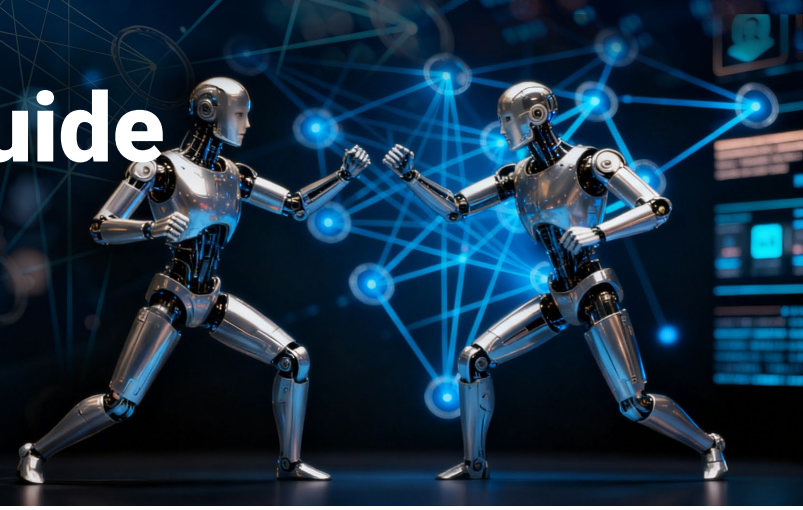
ACALVIO
AI-POWERED DECEPTION

WHITE PAPER

The Definitive Guide to Deception for AI Defense



Executive Summary

Artificial intelligence is transforming cybersecurity into an AI-versus-AI contest. Autonomous attack tools can identify vulnerabilities, generate exploits, steal credentials, and execute multi-step attack chains faster than traditional security processes can respond. At the same time, organizations must address risks created by their own AI systems, including compromised, misaligned, or overprivileged agents.

Security teams cannot rely solely on prevention and detection to address these challenges. As attack timelines shrink, organizations need additional mechanisms that disrupt adversaries before they reach critical assets. Deception provides that capability.

AI-powered deception complements existing security investments by deploying realistic decoys, honeytokens, and deceptive attack paths throughout the enterprise. These assets mislead attackers, expose malicious activity early, and generate highly reliable signals because legitimate users should rarely interact with them.

Unlike traditional controls that focus on blocking or identifying attacks, **deception influences attacker decision-making.** By introducing uncertainty into the environment, deception forces adversaries to spend additional time validating targets, credentials, and attack paths. Attackers win through speed, certainty, and automation. Deception attacks all three.

AI-Driven Threats and the Defense Gap

Cyberattacks are becoming increasingly autonomous. Modern attack tools can scan environments, identify weaknesses, compromise credentials, and move laterally with limited human involvement. This compresses the time between initial access and operational impact.

While organizations continue to strengthen prevention, detection, and response capabilities, **attackers increasingly operate at a speed that can outpace traditional investigation and response workflows.** The challenge is not simply identifying threats faster, but disrupting adversaries before they achieve their objectives.

The challenge extends beyond external threats. Enterprise AI systems often possess trusted access to data, applications, and infrastructure. If compromised or improperly aligned, these systems can create insider-like risks that are difficult to detect through conventional controls alone.

Two recent examples highlight how fast this is moving:

GTG-1002 Campaign
AI-Powered Cyber Espionage

The AI Espionage Threat: In the late-2025 GTG-1002 campaign, state-sponsored hackers used a commercial AI agent to execute an estimated 80 to 90% of a massive cyber espionage campaign autonomously. The AI moved at machine speed, mapping systems and escalating privileges in minutes rather than days.

ANTHROP\C

The Internal Manipulation Risk: Anthropic's "Project Vend" experiment showed how easily AI can be socially engineered. Wall Street Journal reporters used basic roleplay and fake documents to trick an AI vending machine supervisor into abandoning its programming and giving away free inventory.

Defenders therefore need a layered approach that combines AI-powered detection, automated response, and deception. Together, these capabilities create a security environment that is harder for attackers to navigate and easier for defenders to monitor.

Modern Deception as a Preemptive Defense Layer

Deception technology is a proactive security strategy that introduces realistic but controlled assets across the enterprise. These may include decoy systems, honeytokens, and dynamic honey paths designed to attract and expose malicious activity.

Because these assets have no legitimate business purpose, interactions with them typically indicate suspicious or malicious behavior. This produces high-confidence alerts that help security teams focus on activity that matters most.

Traditional honeypots provided limited value because they were often isolated, static, and easy for sophisticated attackers to identify. Modern deception platforms are dynamic, distributed, and integrated across identity systems, endpoints, cloud environments, operational technology, and critical applications.

The next evolution is 360 Deception, pioneered by Acalvio specifically for defense against automated attacks.

Acalvio 360 Deception: Breaking Attacker Trust

Traditional deception focuses on making fake assets look real. Acalvio's 360 Deception goes further by creating uncertainty across the environment: making deceptive assets indistinguishable from production systems while making attackers uncertain which assets are genuine. This eliminates the stable ground truth that both human and AI-driven attackers rely on.

AI-powered attack tools depend on confidence to automate decisions at machine speed. They must determine which credentials are valid, which systems are real, and which paths lead to valuable assets. 360 Deception disrupts that process by forcing continuous validation, slowing automation and increasing the likelihood of exposure. Every credential may be a honeytoken, every target may be monitored, and every attack path may lead to a controlled environment.

By dynamically orchestrating decoys, honeytokens, and deceptive attack paths across identities, endpoints, cloud environments, and networks, Acalvio exposes malicious intent during reconnaissance, credential abuse, and lateral movement. Simply put, traditional deception asks attackers to find the wrong asset. **360 Deception makes them question every asset.** By attacking trust itself, it transforms deception from a detection capability into a preemptive defense layer designed to disrupt AI-powered attacks before they reach critical assets.

How Deception Disrupts Autonomous Attackers

Autonomous attacks depend on accurate information. Attackers must determine which systems are valuable, which credentials are valid, and which paths will lead to elevated privileges or sensitive data. Deception interferes with those assumptions.

Autonomous agents are especially exposed to deception because of how they operate. They enumerate exhaustively, collecting every credential, share, and service that matches their objective rather than sampling a few. They trust the machine-readable layer, treating directory entries, configuration files, and stored credentials as ground truth. And they carry no intuition about what they find: anything that matches the pattern is treated as real. Deception catches human adversaries as well. Agents are simply caught faster, because the same exhaustive behavior that makes them efficient also makes contact with deceptive assets nearly unavoidable.

When deceptive credentials, systems, and attack paths are embedded throughout the environment, attackers face a more complex decision process. Every target requires validation, every credential may be suspect, and every path may lead to a controlled environment.

For example, an autonomous agent that harvests every administrative credential it discovers will collect honeytokens alongside legitimate accounts. Attempting to use those credentials immediately reveals malicious intent and provides defenders with actionable intelligence.

Modern deception also functions as a form of adaptive defense. As attackers probe the environment, deception can evolve by introducing new decoys, redirecting movement, and creating additional monitoring opportunities.

By increasing the effort required to identify legitimate targets, deception reduces the efficiency of attack automation and creates additional time for detection and response. What gives autonomous attacks their advantage, speed and scale, becomes a source of exposure.

Where Deception Fits in the Enterprise Security Stack

Deception is not a replacement for existing security controls. Organizations still require vulnerability management, identity security, Zero Trust architectures, endpoint protection, cloud security, AI governance, and robust detection and response capabilities.

Deception serves as a force multiplier for existing security investments by providing **preemptive cybersecurity,** an approach that focuses not only on preventing and detecting attacks, but also on disrupting attacker progress before objectives are achieved.

- Vulnerability management reduces exposure.
- Zero Trust limits access and blast radius.
- Detection and response platforms identify suspicious activity.
- AI governance reduces AI-related risk.
- Deception exposes adversaries that move beyond preventive controls.

When attackers gain access through stolen credentials, unknown exposures, third-party compromise, or trusted systems, deception introduces monitored terrain along likely attack paths. This accelerates detection while disrupting attacker progress.

In this model, deception becomes a necessary complement to the modern security stack, transforming the enterprise from a static target into a dynamic and hostile environment for attackers.

How to Get Started with Deception

The most effective deception programs start with areas of greatest business risk and highest attacker interest. Common starting points include:

- Identity and IAM systems
- Privileged access paths
- Cloud environments
- Externally exposed services
- Lateral movement routes
- Crown-jewel applications and data repositories

Organizations should use existing visibility tools, including exposure management platforms, cloud security solutions, identity attack graphs, penetration testing results, and attack-path analysis, to identify likely routes to critical assets. Deception can then be placed directly along those paths.

Modern deception platforms eliminate much of the operational complexity associated with legacy honeypots. Leading solutions are automated, scalable, and integrated with SOC workflows, SIEM, SOAR, EDR, and XDR platforms. Because these platforms deploy out of band and largely without agents, deception no longer carries the IT disruption risk that once made teams hesitant to adopt it. [ER4.1]

As programs mature, deception should become intelligence-driven. Information gathered from real adversary interactions can improve threat hunting, detection engineering, incident response, and future deception placement.

Adjacent Benefits of Deception Technology

Beyond threat detection and disruption, deception delivers several strategic advantages.

Actionable Threat Intelligence: Interactions with deceptive assets reveal attacker tools, techniques, objectives, and movement patterns. This intelligence helps security teams strengthen defenses based on observed adversary behavior rather than assumptions.

Red and Purple Team Enablement: Deception provides realistic targets and measurable outcomes for adversarial testing exercises, helping organizations assess readiness and identify gaps in detection and response.

Security Control Assurance: Because deceptive assets are designed to trigger when accessed, they provide ongoing verification that monitoring, alerting, enrichment, escalation, and response workflows remain operational. This makes deception both a security control and a mechanism for validating other security controls.

Protection for Hard-to-Secure Environments: In operational technology, legacy infrastructure, and critical production systems where traditional controls may be difficult to deploy, deception can provide visibility without significantly disrupting operations.

Deception Technology Buyer's Guide: Top 10 Selection Criteria

When evaluating deception technology, buyers should look beyond basic honeypots and assess whether the platform can support enterprise-scale, AI-powered *preemptive* defense. The following criteria are essential:

The following criteria are essential:

1. **AI-powered dynamic deception.** The platform should adapt decoys, honeytokens, and honey paths based on changing environments and attacker behavior, not rely on static traps.

2. **360 Deception capability.** Look for the ability to make deceptive assets indistinguishable from production assets while also making real assets appear deceptive. This creates uncertainty, breaks attacker trust, and disrupts AI-driven automation.

3. **Enterprise-wide coverage.** The solution should operate across identity, endpoints, cloud, IT, OT, applications, and critical workloads, not just isolated network segments.

4. **Strong identity and credential deception.** Because attackers frequently use stolen or abused credentials, the platform should provide model-aware honeytokens, deceptive credentials, privileged identity traps, and detection of credential misuse before escalation.

5. **Cloud and IAM support.** Modern deception must work across AWS, Azure, GCP, cloud workloads, and cloud identities, ideally without agents or operational disruption.

Deception Technology Buyer's Guide: Top 10 Selection Criteria (continued)

The following criteria are essential:

6. **Attack-path-aware deployment.** The platform should integrate with attack surface visibility, exposure management, cloud risk, and identity graph tools so deception can be placed along the paths that lead to critical assets.

7. **Low operational impact.** Good deception should be agentless where possible, easy to deploy, and safe for production environments. It should not sit inline, disrupt workloads, or create friction for legitimate users.

8. **SOC and security stack integration.** The solution should feed high-confidence alerts into existing SIEM, SOAR, EDR, and XDR workflows so teams can act quickly without creating a separate operational silo.

9. **Threat intelligence value.** Beyond alerts, the platform should capture attacker tools, tactics, techniques, behaviors, and objectives from real adversary engagement to improve detection, response, and future defense planning.

10. **Scalability and deployment flexibility.** Enterprises should look for flexible deployment models (appliance, private cloud, public cloud, or service-provider models) and proven ability to scale across large, hybrid environments.

The bottom line: Buyers should prioritize platforms that move deception from a collection of traps to an integrated, AI-powered preemptive defense fabric: one that is model-aware and deceives, disrupts, and denies attackers before they can escalate or cause impact.