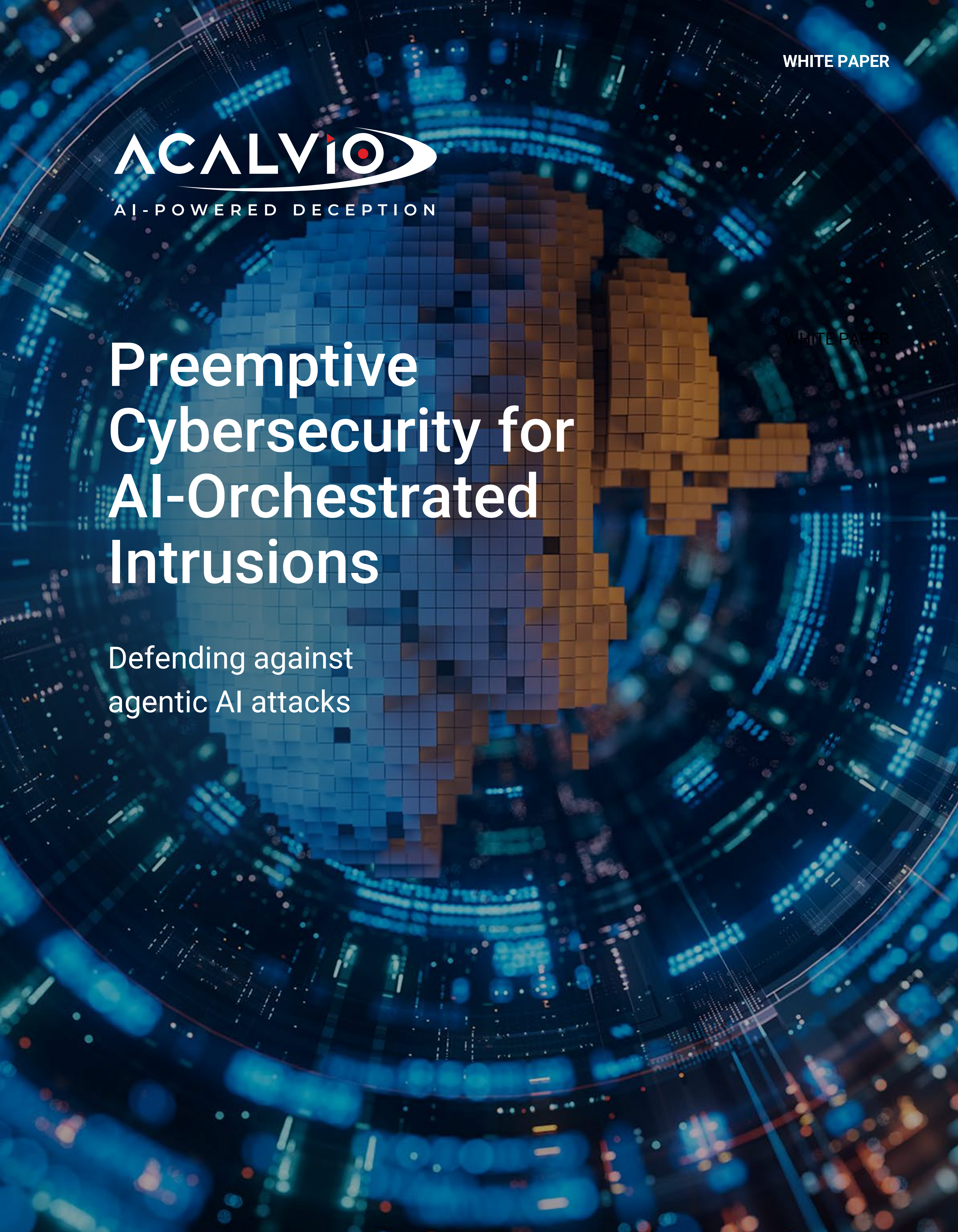
WHITE PAPER

ACALVIO

AI-POWERED DECEPTION



Preemptive Cybersecurity for AI-Orchestrated Intrusions

Defending against agentic AI attacks

WHITE PAPER

What is an AI-orchestrated attack?

AI-orchestrated attacks involve attackers using AI agents to autonomously plan, execute, and adapt multi-stage intrusions. In these exploits, the human attacker provides a target or objective as the initial prompt to the AI agent. The agent uses AI to autonomously execute the full attack lifecycle, from reconnaissance to exfiltration. The agent breaks down the exploit into subtasks, with each task being executed in an autonomous manner.

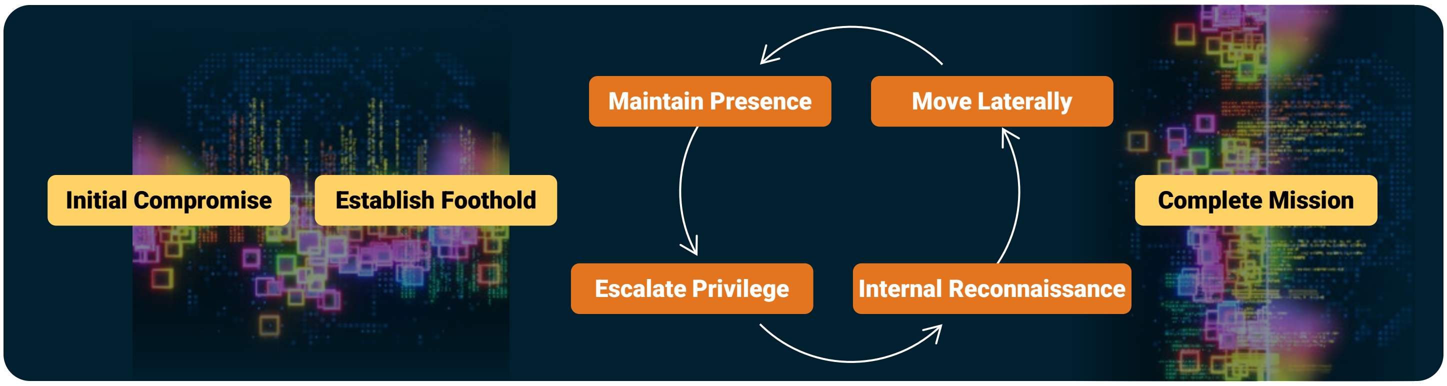


Figure 1: Threat actors are leveraging AI across all stages of the attack lifecycle

Source: https://cloud.google.com/blog/topics/threat-intelligence/distillation-experimentation-integration-ai-adversarial-use

Examples of AI-orchestrated attacks: Anthropic published a report that highlighted a multi-stage attack campaign orchestrated by AI. The AI agent performed a multi-stage exploit, from reconnaissance to data exfiltration, with minimal human intervention. This exploit, orchestrated by a threat actor group code named GTG-1002, exploited known vulnerabilities at machine-speed, achieving success without requiring the use of zero days.

A second example: Sysdig threat research published a detailed report on an autonomous cloud intrusion, an exploit called LLMJacking. The exploit started with an AI agent finding credentials in a public S3 bucket, performing identity exploits to gain administrative privileges in the cloud, and then launching foundation models.

How are AI-orchestrated attacks different?

Cyber attacks have required skill and sophistication from the threat actor. Performing advanced exploits, such as living off the land attacks where the attacker uses built-in tools and utilities on an endpoint, requires deep knowledge of operating systems, network infrastructure, and offensive security. This has represented a high bar and has limited these exploits to a set of APT threat actor groups and adversaries with advanced skillsets.

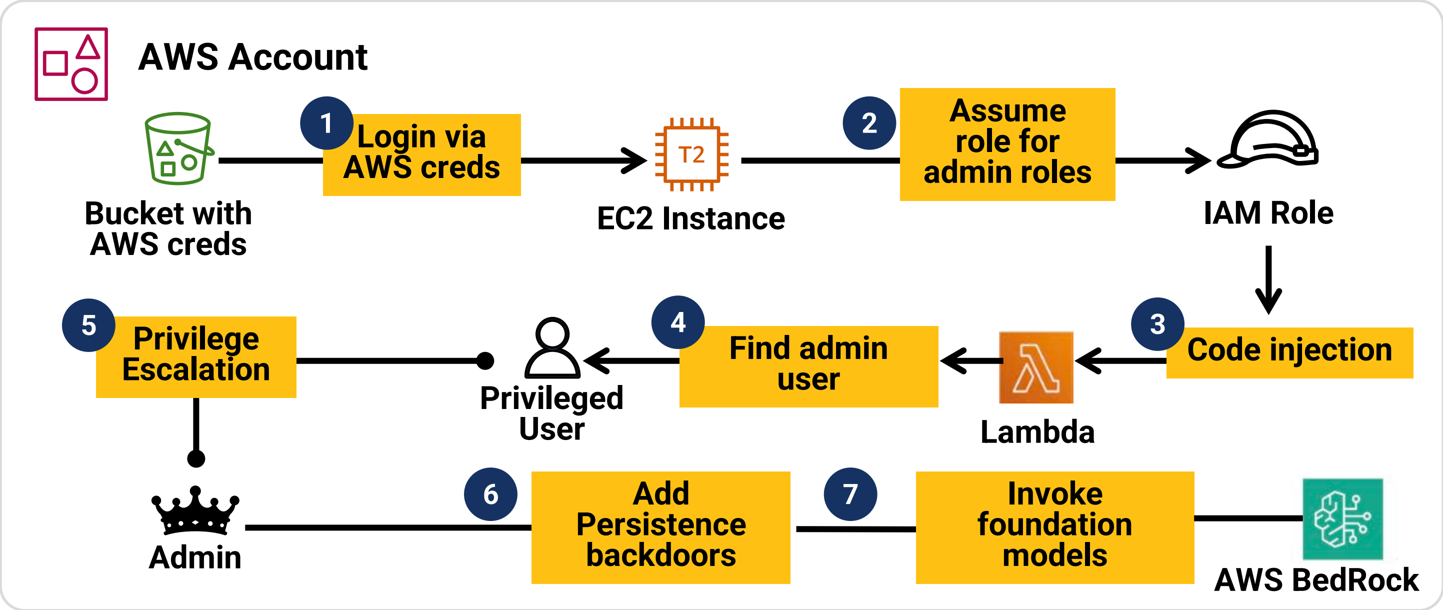


Figure 2: LLM assisted multi-stage exploit

How are AI-orchestrated attacks different? (cont.)

Although these techniques are well known and well researched, the skill required to execute them successfully has limited their use. Attacker groups that have the requisite skills are limited by human processing capabilities, with each step of an exploit sequence requiring human cognition and processing. Attacker dwell time has historically averaged days.

The advances in AI are transforming the threat landscape. As highlighted in the exploits above, AI agents are capable of performing autonomous exploits, including multi-stage exploit sequences. These exploits are occurring at machine speed, reducing the elapsed time windows by an order of magnitude. In the cloud intrusion example, the attacker gained administrative privileges in the cloud in 8 minutes, representing a shift not previously observed at that speed. AI agents are capable of generating novel exploit combinations for each exploit execution cycle, marking another departure from human-driven exploits.

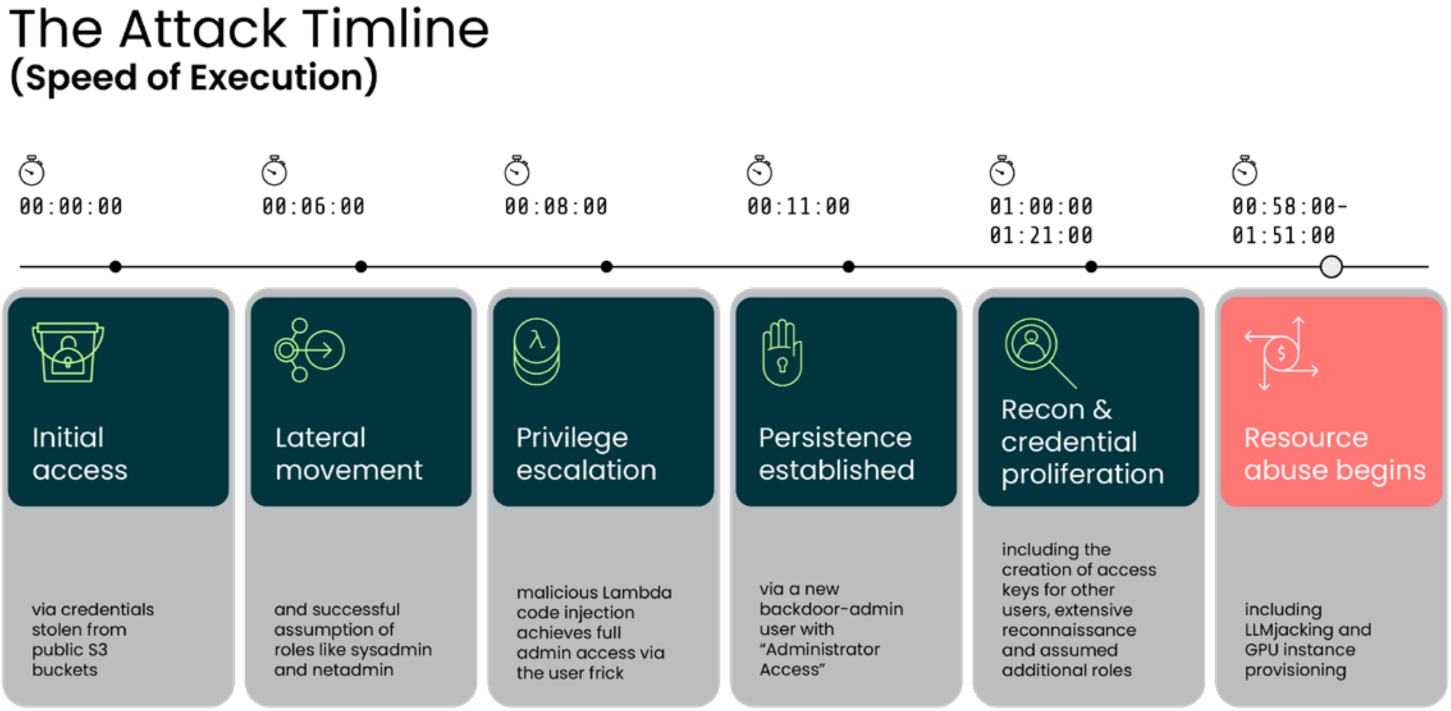

The Attack Timline
(Speed of Execution)
00:00:00
00:06:00
00:08:00
00:11:00
01:00:00
01:21:00
00:58:00-
01:51:00
Initial access
via credentials stolen from public S3 buckets
Lateral movement
and successful assumption of roles like sysadmin and netadmin
Privilege escalation
malicious Lambda code injection achieves full admin access via the user frick
Persistence established
via a new backdoor-admin user with "Administrator Access"
Recon & credential proliferation
including the creation of access keys for other users, extensive reconnaissance and assumed additional roles
Resource abuse begins
including LLMjacking and GPU instance provisioning

Figure 3: Attack timeline of AI agent-based cloud intrusion

Source: https://www.sysdig.com/blog/ai-assisted-cloud-intrusion-achieves-admin-access-in-8-minutes

What are the implications on cyber defense?

The footprint of AI agents in enterprises has rapidly increased with the adoption of AI into enterprise workflows. Attackers do not need to bring in new agents into the organization; the existing agents are manipulated into offensive actions through carefully crafted prompts. This implies that the attacker can always gain the initial beachhead into the organization. From a defender's perspective, an assume compromise security posture is prudent given this threat landscape.

Defenders have traditionally deployed reactive threat detection systems to detect and respond to malicious activity. These systems, leveraging principles of anomaly detection or rule-based analytics, are based on observing activity in the environment and looking for patterns that indicate malicious activity. Reactive security requires a sufficient signal before flagging as suspect, which requires an extended time window for alert identification.

The detection time windows of reactive security controls far exceeds the machine-speed at which AI-orchestrated attacks are executed, making it possible for agentic exploits to successfully complete their offensive campaigns.

AI agents are capable of planning exploits over long horizons of time, that exceed human capacity. By planning low and slow reconnaissance that stays within normal usage patterns, and then rapidly executing the remaining steps of the multi-step exploit campaign at machine-speed, agents can initially remain undetected and then overwhelm the defender that has to perform human correlation of alerts to identify malicious activity.

Reactive security solutions based on traditional detect and respond also have a security gap related to exploits targeting vulnerabilities in AI infrastructure itself. AI Agents, MCP servers introduce new vulnerabilities in both application and infrastructure layers. These are being exploited by threat actors. The existing tooling is unaware of the characteristics of this infrastructure, resulting in an expanded detection gap and exploitable attack surface.

To combat these attacks, defenders need a shift toward proactive and preemptive cybersecurity controls that anticipate attacks and detect these early in the exploit lifecycle. The approach protects the existing assets in the organization and also safeguards the AI infrastructure itself from targeted exploits.

What is preemptive cybersecurity and what role does it play in combating AI-orchestrated attacks?

The Anthropic report discusses the countermeasures for the autonomous exploit sequence of AI agents, highlighting the need for **“proactive and early detection systems”** to combat autonomous cyber attacks. Gartner has published research on these exploits, calling for a shift from reactive security to preemptive defense.

Preemptive cybersecurity is based on the premise of anticipating attacks and setting proactive traps for AI-orchestrated attacks. These traps (in the form of deceptions) are designed to detect threats at the early stages of the exploit lifecycle, typically at the reconnaissance or credential access stage. Defenders place traps at strategic locations of interest to the agent, typically near high-value assets and on attack pathways leading to the assets. Any activity involving these traps is an immediate indicator of malicious intent, greatly reducing the detection time window. By identifying the agentic AI exploit at the reconnaissance stage itself, defenders gain the ability to orchestrate response actions to neutralize the exploit sequence and protect high-value assets.

In addition to early detection, deception serves as an instrument to **shape shift** the AI agent’s perceived reality. This is achieved through the principles of **divert and deflect**, an approach designed to disrupt the AI agent’s exploit pathways

Example scenario: preemptive defense to combat the AI-orchestrated cloud intrusion

Let us analyze the role of preemptive defense to combat AI-orchestrated attacks. The defender places traps for the AI agent, these traps are in the form of decoys (deceptive cloud resources) and honeytokens (deceptive identities and data). The traps are placed at strategic locations of interest to the AI agent, these include IAM stores, credential stores such as cloud key vaults, high-value assets such as cloud databases and cloud services providing AI infrastructure (such as AWS Bedrock).

The AI agent gains initial access and attempts the exploit sequence. As the AI agent enumerates identities and attempts to exploit these to elevate privileges, the agent encounters honeytokens that represent deceptive identities of interest. The agent attempts to impersonate this identity, raising an actionable alert. The placement of a tailored combination of honeytokens and decoys results in the agent finding and pursuing alternate pathways that lead the agent away from the original goal or mission.

The defensive landscape is dynamically updated, resulting in an unpredictable environment that introduces uncertainty for the AI agent. By controlling the agent's reality, defenders gain an effective approach to combat AI-orchestrated attacks.

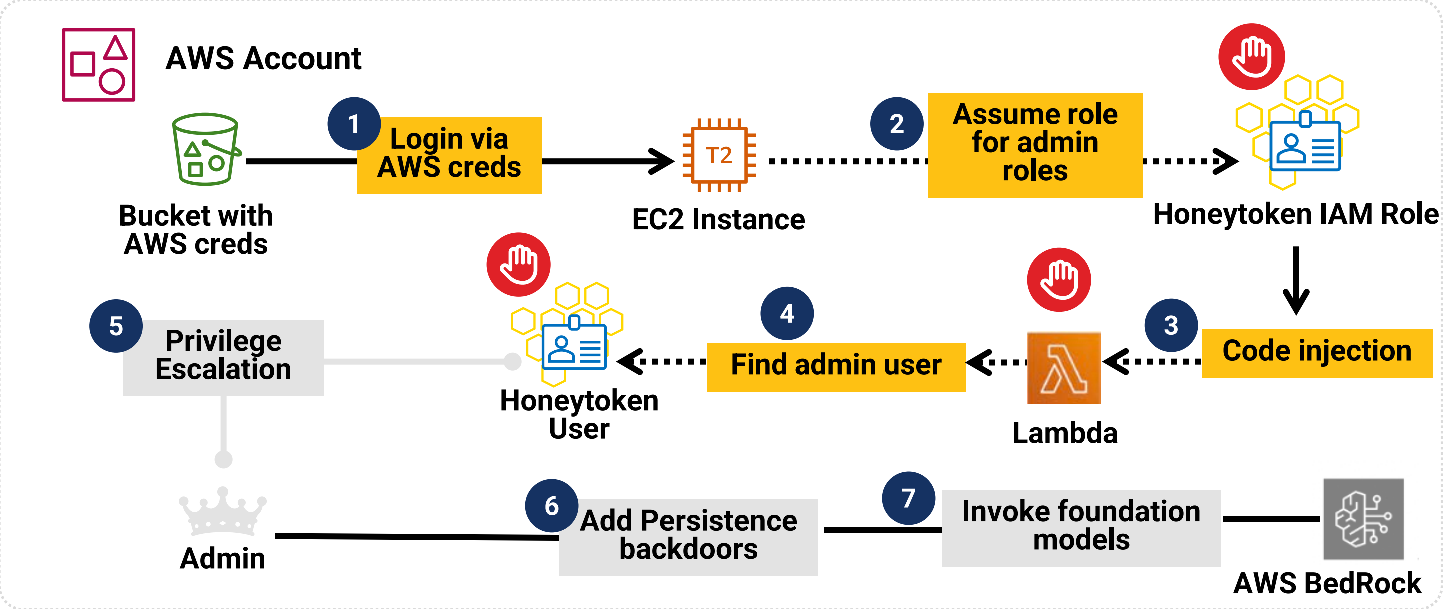


Figure 4: Deception-based preemptive defense stops this exploit at step 2

How does Acalvio ShadowPlex help operationalize this strategy?

Security teams transitioning to AI-driven cyberdefense must move beyond manual configuration to solve a fundamental strategic puzzle: how to outmaneuver an adversary that processes information at machine speed. The challenge lies in creating a precisely balanced automated defense, calculating the exact density and placement of digital tripwires so that an attacker, no matter how sophisticated, is statistically guaranteed to engage with a decoy rather than a real asset. At enterprise scale, the number of possible attack paths is too vast for human logic to map -- it requires an AI-orchestrator capable of seeing the entire digital chessboard and adjusting the pieces in real-time.

Acalvio ShadowPlex serves as a preemptive, autonomous security engine that treats the design of these digital tripwires as a continuous optimization problem. By integrating strategic logic with AI-driven action, ShadowPlex dynamically evolves the defensive perimeter to stay ahead of automated, self-learning threats. In cloud-native environments, ShadowPlex architecturally embeds agent-aware deceptions, deploying synthesized honeytokens (IAM identities) and decoys (compute/storage) specifically designed to disrupt the model-based logic of agentic AI exploits. This transforms the cloud into a hall of mirrors where an attacker’s automated tools are systematically led toward paths of maximum futility.

Acalvio is the leader in autonomous cyber deception, defending against APTs, insider threats, and ransomware. Its AI-powered Preemptive Cybersecurity Platform, protected by 25 patents, delivers threat detection across IT, OT, and cloud environments and advances Identity Threat Detection and Response (ITDR) with Honeytoken-driven Zero Trust security.

Based in Silicon Valley, Acalvio serves midsize to Fortune 500 companies and government agencies, offering flexible deployment from the Cloud, on-premises, or through managed service providers. **www.acalvio.com**

© 2026 Acalvio Technologies, Inc. All rights reserved.