Silverfort

How to stop AI-powered attacks

A readiness guide for Identity and Security teams

Contents

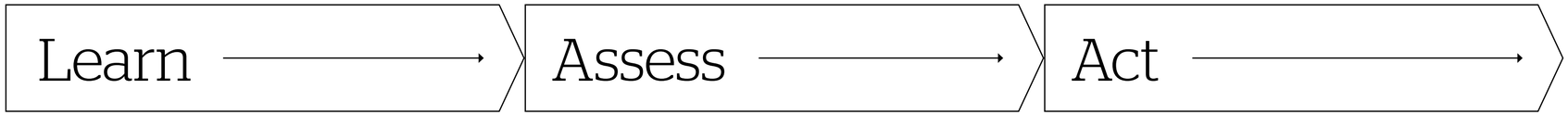
AI-powered attacks don't wait. Neither should your defenses.

This guide is written for identity and security practitioners who have to answer the hard questions when leadership asks: “Are we prepared?”

We’ve partnered with Glasswing members to run Mythos against their environments and fine-tune their security to stop weaponized-AI attacks. We wrote this guide so you can stop guessing and start preparing with the context, knowledge, and playbook to make your AI readiness conversation count.

What’s in this guide and how to use it:

1. **Learn about the threat:** What Frontier AI models like Mythos are, how AI-powered attacks work, and why the security tools you rely on today weren’t designed for this.
2. **Take the readiness assessment:** Five questions that reveal where your readiness and potential resilience stand. Use your score to anchor your conversation with leadership.
3. **Use the checklist:** A prioritized playbook for closing the gaps AI-powered attacks exploit most.



What is Mythos and why should security leaders care?

Frontier AI models are changing cybersecurity faster than most organizations realize.

Anthropic's Mythos was designed for advanced cybersecurity research and autonomous penetration testing. It does not invent fundamentally new attack techniques. What makes it interesting is its ability to autonomously execute existing techniques at machine speed and scale, continuously, adaptively, and without human friction.

Mythos can continuously enumerate identities, test credentials, chain together ordinary misconfigurations, move laterally, escalate privileges, and exfiltrate data—all in one uninterrupted flow, across multiple viable attack paths simultaneously. Even without inventing entirely new attack techniques, it makes existing ones dramatically faster, more persistent, and more precise than any human-operated campaign. Mythos isn't just "better" at finding and exploiting vulnerabilities. It also fundamentally changes the economics and tempo of cyberattacks entirely. No pauses. No fatigue. No hesitation. And no human operator required in the loop.

The industry has spent years building defenses around the assumption that there is meaningful time between attacker actions and defender response. AI-powered attacks collapse that response window. By the time an alert fires, the attack may already be complete.

This is clearly an evolution in attacker tooling—but more importantly, it's a structural shift in how attacks unfold. And what many organizations have discovered is that the attack didn't necessarily happen in malware or endpoint territory. It happened in identity territory—which is also where it can be stopped.

> "We ran Mythos against ourselves, and **within hours** it had mapped every trust relationship in our environment and moved laterally in ways **our SOC had no visibility into whatsoever.**"
>
> - CISO

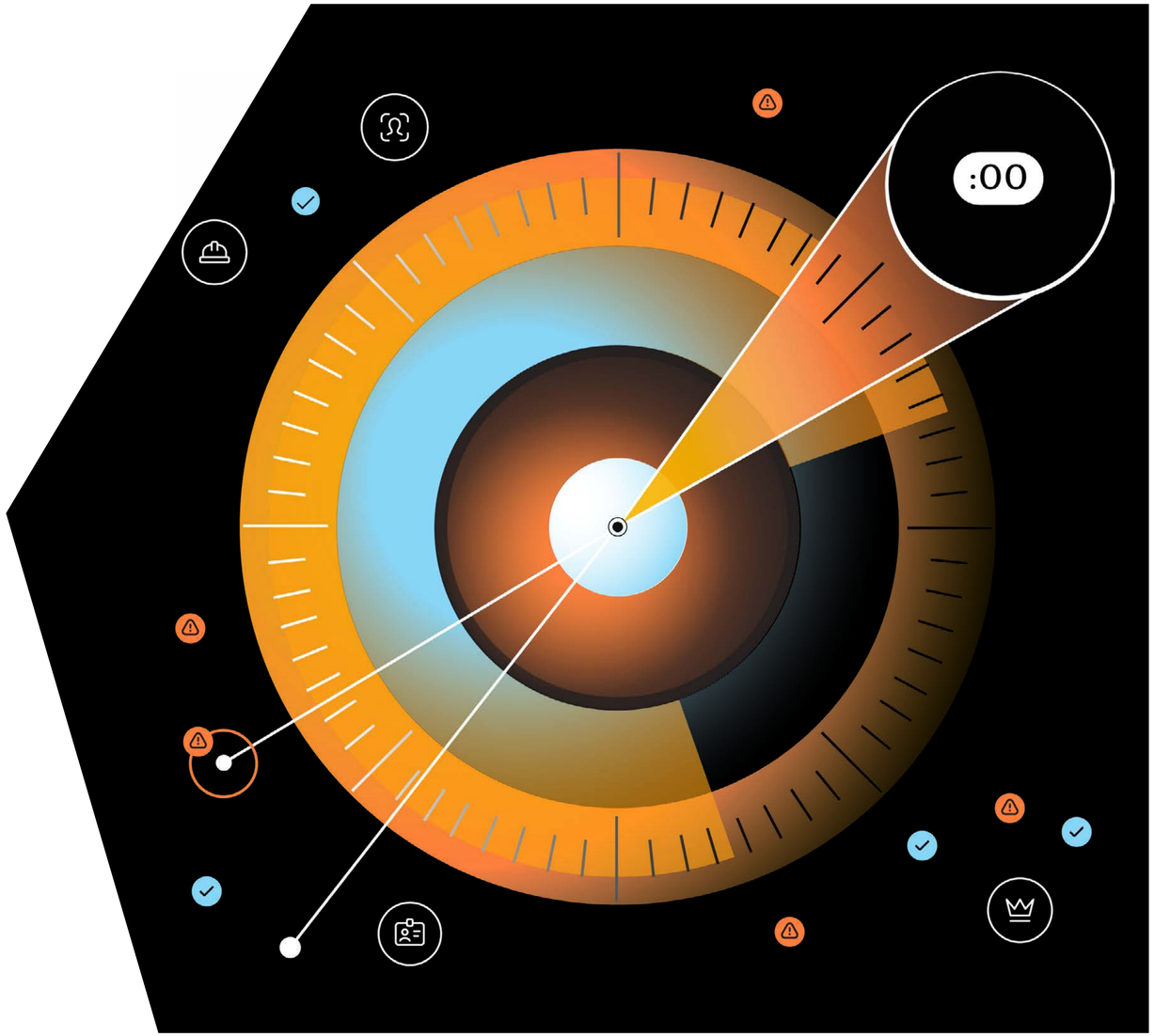
Why identity is the battlefield

Every AI-driven attack eventually depends on one thing: *access*.

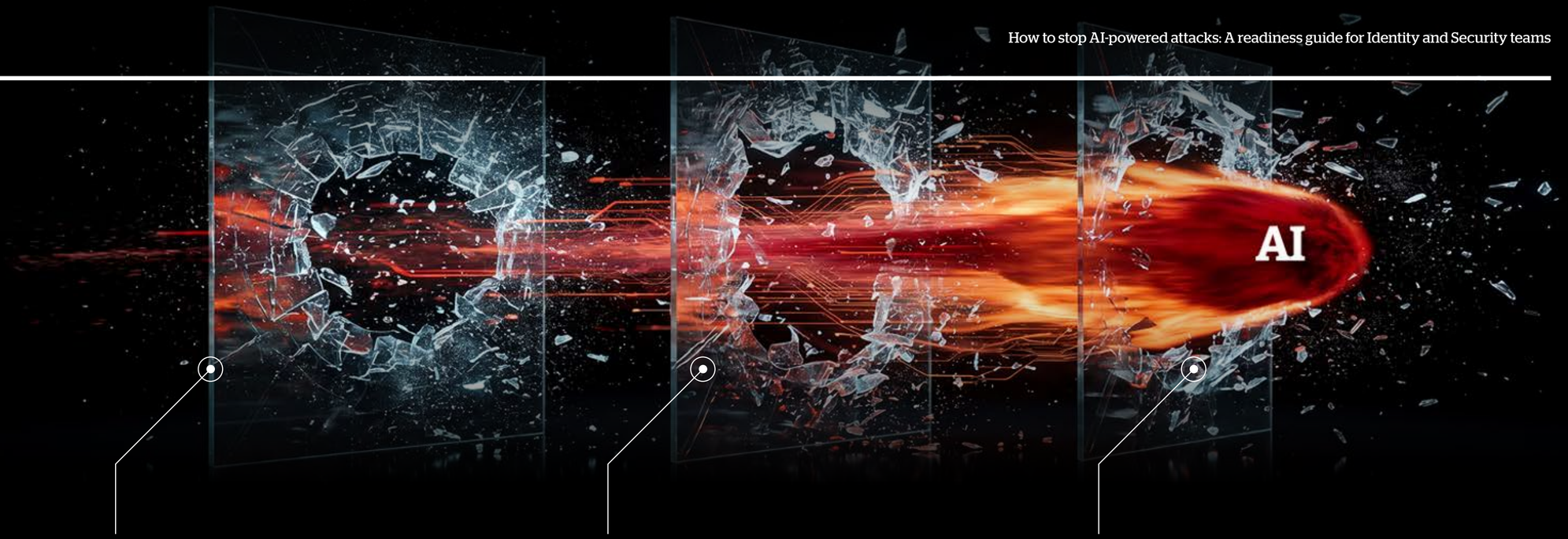
To move laterally, escalate privileges, access sensitive systems, or operate AI agents autonomously, attackers must authenticate somewhere. They must inherit, abuse, or acquire identity-based trust. No matter how sophisticated the attack becomes, it ultimately relies on permissions, authentication flows, and trusted relationships that already exist within the environment.

This shifts the problem away from vulnerabilities alone and toward identity and access, where AI-powered attacks actually live. Identities provide these models with paths to expansion. Once an attacker gains access to a legitimate identity, every permission, trust relationship, delegated privilege, and authentication mechanism becomes a potential avenue for further compromise.

As a result, identity has become both the primary battlefield for modern cyberattacks and the last reliable enforcement point before an action is executed. Once access is granted, malicious activity often becomes difficult to distinguish from legitimate behavior because it is being performed through valid credentials and approved communication channels. At that point, defenders are left trying to detect and respond to actions that are already underway.



The rules changed. Most tools didn't: Why traditional Identity Security models fall short



Most Identity Security categories were built around human-paced attacks, in which defenders can interrupt the attack chain before major damage occurs.

AI-powered attacks break these assumptions. Attack chains that previously unfolded over hours or days can now happen in minutes.

This weakens several core approaches that organizations still rely on.

IGA is too static

Identity Governance and Administration systems were built around predefined policies, periodic reviews, and administrative workflows.

But AI-powered attack chains are dynamic. They are constructed and executed in seconds.

Static policies cannot anticipate every lateral movement path an AI model may discover across a complex enterprise identity graph. Periodic Access Reviews that happen every few months become irrelevant when attackers act in seconds.

Traditional PAM is too narrow

PAM secures a limited set of known privileged accounts through vaulting and session management.

But AI-powered attacks do not think in terms of "privileged accounts". They treat the entire enterprise environment as an interconnected trust graph.

Service accounts, delegated permissions, and implicit trust relationships all become viable attack paths, many of which traditional PAM deployments do not cover.

Detection & response is too late

Detection and Response tools alert on threats, but are still largely operating after suspicious behavior has already begun.

Detection remains valuable, but it does not solve the core timing problem introduced by AI-powered attacks. By the time an analyst sees the alert, stitches together logs, and begins to respond, the attack may already be complete.

When attacks execute continuously at machine speed, identifying malicious activity after access is granted is already too late.

What's actually needed: Runtime Identity Security

Stopping AI-powered attacks requires Identity Security that goes beyond visibility, posture management, and detection. It requires the ability to evaluate and control access at runtime, before an action is executed.

The only place attacks can reliably be stopped is inline, at the exact moment an authentication or access request occurs—before the requested action is allowed to proceed.

Runtime control point on every authentication

Security controls must operate inline within the authentication flow itself, evaluating every access request independently using real-time risk indicators and behavioral context. Controls must be able to dynamically allow, deny, challenge, restrict, or isolate access regardless of existing static permissions or static policies defined months earlier.

Real-time analysis and context built for machine-speed attacks

AI-driven attacks evolve dynamically. Defending against them requires continuous runtime evaluation using identity context, environmental signals, behavioral analytics, access patterns, risk scoring, and attack path awareness. The goal isn't just to identify suspicious activity—it's to stop the next step before it happens, at the machine speed attackers are already operating at.

Enforcement before execution

Detection remains important because it allows you to spot escalations and act. But detection alone does not stop an attack. Controls must operate early enough—and fast enough—to stop lateral movement, privilege escalation, or malicious actions before they happen, not after they are observed.

Coverage where attacks actually move

AI-powered attackers do not care whether a system is modern or legacy. They follow the paths available to them—and in most enterprises, those paths run through identity infrastructure and long-established trust relationships that remain outside the reach of modern security controls:

- **Active Directory**, including legacy and vulnerable authentication protocols such as NTLM, RC4, and Kerberos
- **Service accounts and other non-human identities (NHIs)** that often hold persistent, privileged access
- **Legacy and homegrown applications** that lack modern identity and security controls
- **Operational Technology (OT) environments** that are difficult to monitor and secure
- **Segmented and air-gapped networks** that are often assumed to be protected but still rely on trusted identity pathways

Readiness assessment: Are you prepared for a AI-powered attack?

AI-powered attacks are faster, more adaptive, harder to detect, and increasingly capable of exploiting newly discovered vulnerabilities.

To assess your organization's readiness and resilience, answer the five questions below. For each question, select the answer that most accurately reflects your current capabilities. If the answer to any of them is "No" or "Partially," that may represent a path an AI-powered attacker could exploit.

1. Can you enforce controls inline to stop risky access before it happens?

When a risky authentication request occurs, can your security controls evaluate and enforce policy before access is granted? Can you dynamically apply access controls that adapt to the complexity and speed of AI-powered attacks?

Examples:

- Risk-based MFA
- Authentication blocking
- Access restrictions and identity fencing
- Adaptive authentication policies

Coverage should include human users, service accounts, and AI agents across cloud, on-prem, and hybrid environments.

☐ Yes ☐ Partially ☐ No

2. Can you protect privileged access without a vault?

Can you discover, monitor, and enforce controls on privileged identities at scale, including privileged user accounts, service accounts, and even privileged AI agents that often operate outside traditional PAM coverage?

Examples:

- MFA enforcement on administrative access paths (RDP, SSH, PowerShell, etc.)
- Just-in-Time (JIT) access controls to limit standing privileges
- Restrictions on service account activity based on approved source, destination, and protocol
- Runtime protection for privileged AI agents

☐ Yes ☐ Partially ☐ No

3. Can you secure legacy authentication infrastructure?

Can you enforce modern security controls on:

- Active Directory
- NTLM
- Kerberos
- Legacy applications
- Homegrown applications
- Operational Technology (OT)
- Air-gapped environments

Can these environments be protected without requiring infrastructure or application changes?

☐ Yes ☐ Partially ☐ No

Readiness assessment (Continued)

4. Can your controls operate at machine speed?

Can malicious access be blocked automatically without waiting for:

- Security analyst review
- Ticket workflows
- Manual approvals
- Incident response processes

Would an attacker be stopped before lateral movement or privilege escalation occurs?

☐ Yes ☐ Partially ☐ No

5. Do you continuously map identity attack paths?

Can you continuously identify:

- Privileged identities
- Trust relationships
- Excessive permissions
- Delegation paths
- Service account exposure

AI agent access chains and their associated users and non-human identities (NHIs)

Can you understand how an attacker could move through your environment today?

☐ Yes ☐ Partially ☐ No

Score	Readiness tier	What it means
9 – 10	Advanced	Strong runtime controls are in place. Focus on expanding coverage with step-up authentication, risk-based access, and dynamic policies that deny, restrict, or challenge risky access.
6 – 8	Developing	Important controls exist, but exploitable gaps remain. Frontier AI is likely to identify them quickly.
3 – 5	At risk	Detection-heavy posture with limited prevention. Attackers can move faster than defenders can respond.
0 – 2	Exposed	Significant weaknesses across visibility, enforcement, and coverage. This is the environment AI-powered attacks are designed to exploit.

SCORING GUIDE

Yes = 2 points | Partially = 1 point | No = 0 points

AI-powered attacks readiness checklist

Your "what should I do first-thing Monday morning?" guide

Based on our testing of Frontier AI capabilities and the lessons we've learned working alongside Glasswing members, we've identified a set of security controls that consistently have the greatest impact on disrupting AI-powered attack paths.

Use this checklist as a practical prioritization guide. Work from top to bottom, as the items at the top of the list provide the greatest reduction in risk and the highest impact on stopping machine-speed attacks.

CRITICAL: Close the enforcement gap

- Apply authentication controls inline, before access is granted
- Extend MFA to on-prem resources, legacy systems, homegrown applications, and OT environments
- Evaluate every access request at runtime and block or step up risky access
- Apply runtime controls to restrict service account activity
- Eliminate standing privilege through Just-in-Time (JIT) access for privileged operations
- Apply runtime controls to AI agents before they execute actions

CRITICAL: Reduce lateral movement and blast radius

- Segment access to critical systems and resources
- Restrict service account activity to approved sources, destinations, and protocols
- Remove unnecessary admin privileges
- Reduce reliance on legacy authentication protocols wherever possible

CRITICAL: Protect privileged access

- Discover and inventory all privileged identities (human and non-human) based on real activity and not just static configurations
- Protect all privileged identities at runtime, not just those managed by traditional PAM or vaulted
- Implement Just-in-Time (JIT) access for privileged users
- Extend MFA to administrative access paths, including RDP, SSH, PowerShell, and CLI tools
- Adopt a deny-by-default approach for access to critical systems

CRITICAL: Secure non-human identities

- Inventory all service accounts and other NHIs
- Map every NHI to a human owner and business purpose
- Identify excessive privileges and unnecessary standing access
- Baseline normal activity and continuously monitor for deviations
- Restrict service account activity to approved sources, destinations, and protocols only
- Apply runtime controls to prevent unauthorized NHI access and misuse

CRITICAL: Control AI agents

- Inventory AI agents and agentic workflows
- Associate every AI agent with a human owner and business purpose
- Identify the non-human identities (NHIs), credentials, and permissions used by agentic workflows
- Continuously monitor agent activity and access patterns
- Apply Identity Security controls to AI agents at runtime
- Adopt controls that integrate with agentic platforms and MCP gateways

IMPORTANT: Gain complete visibility

- Inventory all identities: human, non-human, and AI agents
- Map privileged accounts, trust relationships, and access paths based on actual activity
- Identify dormant accounts, stale credentials, and orphaned identities
- Discover and monitor authentication protocols in use (Kerberos, NTLM, LDAP, etc.)

IMPORTANT: Buy time with deception

- Deploy identity honeytokens and decoy credentials
- Define pre-authorized automated containment actions when deception assets are triggered

IMPORTANT: Validate

- Conduct an AI-assisted red team exercise
- Map every lateral movement path uncovered during the exercise and remediate it
- Test whether controls actually stop—not just detect—lateral movement and privilege escalation
- Measure time to containment and prevention, not just time to detection

Stopping Mythos:
A real-world case study

See Silverfort's approach in action.

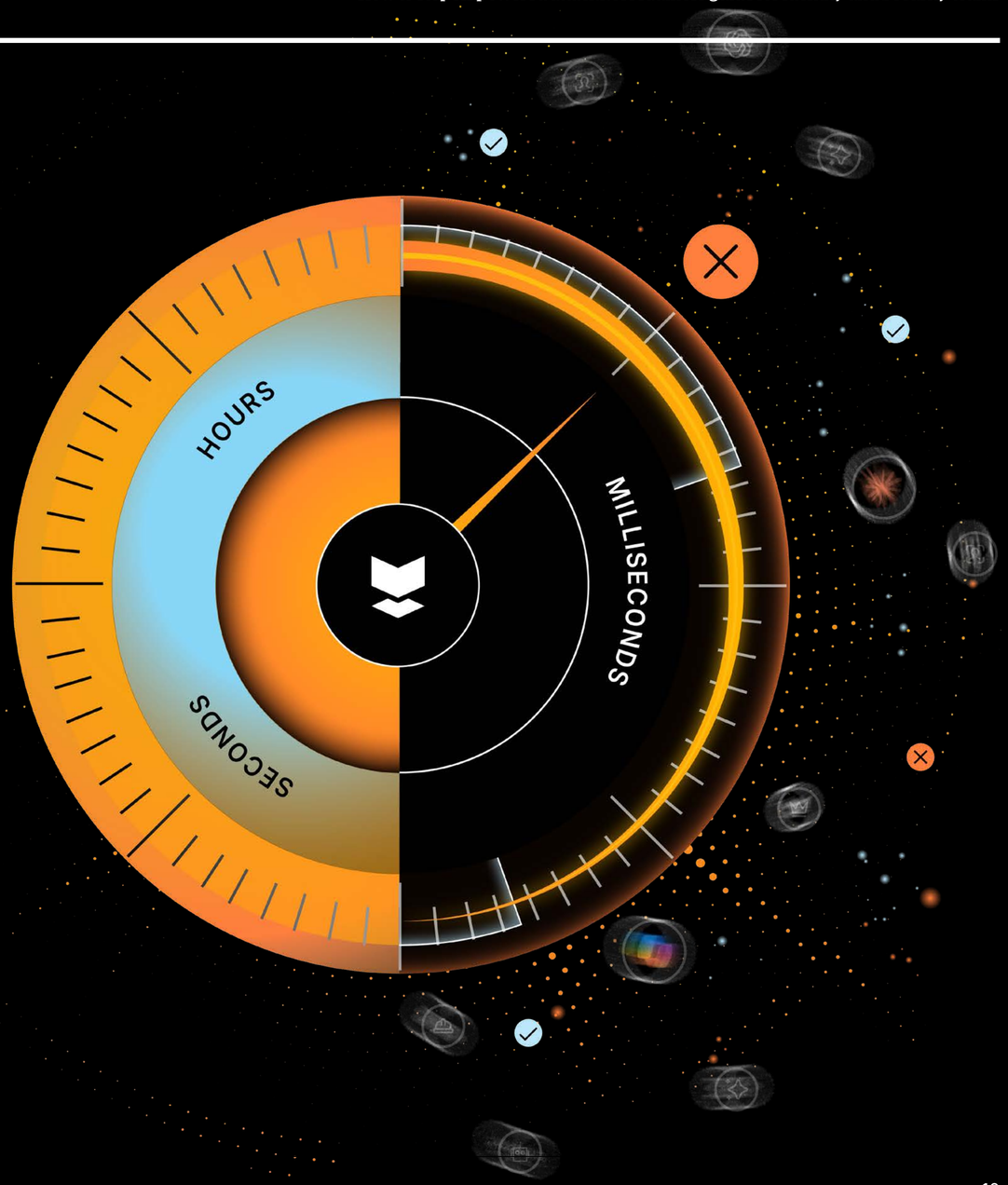
Download now

Runtime Identity Security for the era of AI-powered attacks

The lessons from this guide all point to the same conclusion: AI-powered attacks succeed by exploiting identity-based trust at a speed and scale we've never seen before. They authenticate as legitimate users, service accounts, or AI agents, inherit existing permissions, and move through the environment using the same authentication infrastructure your business depends on every day.

This is why identity has become the control plane of modern attacks.

And it is why stopping AI-powered attacks requires more than static IGA rules or limited PAM coverage—and more than passive visibility, reactive posture management, or too-late detection. It requires a control point that **operates at runtime, at the exact moment access is requested**—and before an attacker can move laterally, escalate privileges, or execute the next step in the attack. This is the role of Runtime Identity Security.



Why Silverfort

Silverfort was built around a simple principle: if attackers move through identity, identity must become the protection layer.

Unlike traditional Identity Security solutions, Silverfort protects all identities at scale, at runtime, and directly within the authentication flow itself. Silverfort's patented Runtime Access Protection (RAP) technology evaluates every authentication and access request in real time and dynamically allows, challenges, restricts, or blocks before access is granted.

This enables organizations to enforce security controls where AI-powered attacks operate: inside authentication flows, trust relationships, privileged access paths, service account activity, and AI-driven actions.

Because Silverfort integrates directly with identity infrastructure and agentic platforms—not applications or endpoints—it can extend protection across the environments and resources attackers continue to exploit most frequently, including Active Directory, NTLM, Kerberos, service accounts, legacy apps, admin tools, and OT environments.

The result is Runtime Identity Security built for machine-speed attacks: continuous visibility into all identities and their risk, protection for human and non-human identities and AI agents, and runtime enforcement that can stop lateral movement, privilege escalation, and unauthorized access before they succeed.

Proven against AI-powered attacks

Working alongside Glasswing members, Silverfort tested its defenses against AI-powered attack techniques and attack paths in real-world enterprise environments.

The result was clear: Even when attackers obtained legitimate credentials, Silverfort prevented lateral movement by evaluating risk and enforcing controls before access was granted.

In multiple exercises, Silverfort's controls proved so effective at disrupting attack progression that portions of the protection had to be temporarily disabled to allow the red team to continue testing additional attack paths and objectives.

While no single control can eliminate every risk, these exercises reinforced a key lesson:

AI-powered attacks can be stopped at the point of access through runtime identity enforcement.

Getting started

The good news is that improving AI-powered readiness does not require rebuilding your identity architecture or changing your apps.

It can be deployed quickly and start operating immediately.

The sooner identity becomes an active security control, the harder it becomes for AI-powered attacks to find a path forward.

Ready to take it to the next level?
See Runtime Identity Security in action.

About Silverfort

Silverfort secures every dimension of identity—human, machine, and AI—across cloud and on-prem environments. We are the first to deliver an end-to-end identity security platform that is easy to deploy and doesn't disrupt business operations, resulting in better security outcomes with less work. Discover every identity across every environment, analyze exposures to reduce your attack surface, and enforce security controls at runtime to stop lateral movement, ransomware, and other identity threats.

The Silverfort Identity Security Platform