semperis

WHITE PAPER

Introducing AI Agents to Your Identity Fabric

Think of AI agents as curious, genius, sociopathic 5-year-olds.

They will explore every corner of your hybrid identity system, looking for ways to get in and take control.

To protect your business-critical functions, you must understand how AI agents work, add risk, and exploit vulnerabilities—and how they can help you prepare and defend your identity fabric for the agentic future.

By **Alex Weinert**
Chief Product Officer, Semperis

AI is the stuff of productivity dreams—and security nightmares.

Agentic AI (agents based on generative AI) has captured the imaginations—and budgets—of organizations around the world.

Today, users are pointing GenAI tools at company resources to output web pages, documentation, emails, and code, while agents accelerate tasks and workflows with autonomous decision making and problem solving, taking actions without human intervention. For organizational leaders, the resulting opportunities for productivity optimization are intoxicating—but for security teams, the arrival of GenAI is terrifying.

This is because **GenAI** *decreases* predictability while it *increases* discoverability of security flaws in your environment. **Agentic AI** tools increase the identities you need to manage and create new challenges for securing the APIs, applications, and services they consume.

The deployment of GenAI tools and agents, whether by your users or adversaries, accelerates the potential for exploiting every vulnerability and permissions gap in your identity fabric —at machine speed.

Gone are the days when you worried that a user would click a phishing link, expose access to their email, and open the path to compromise one account, then move laterally towards your domain controller. Today, users are installing OpenClaw and giving their agents (or the person that hacked their desktop) unfettered access to your environment, enabling blindingly fast, autonomous compromise and attack.

Agentic is coming, and the agents will need very strong boundaries. Your identity fabric is what creates and enforces those safe boundaries for agents, and maintaining its integrity is critical to containing them.

At the same time, productive use of agents requires that those agents have identity and can interact with your identity fabric to get permission to access the key resources that make them so valuable.

Thus, as your business increases its bet on agentic, so too it increases its bet on your identity fabric—because if identity goes down, all your agents do too, along with the business processes they support.

Happily, you're an identity expert, and this is right in your wheelhouse. Enabling productivity while maintaining security is the core domain of identity and access management.

The reality of how GenAI tools work in relation to identity infrastructure—and how that impacts your security and resilience landscape—is the focus of this paper. We'll explore how generative and agentic technologies work and the implications they have for your identity fabric. We'll also dig into how you use that fabric to regulate agents, and what you must do to protect the fabric from malicious agentic use.

Author's note: This is an entirely human-written paper. I used a generative AI assistant for fact checking and research.

Contents

How Are Agents Operating in Your Environment?

Let's start by taking a look at how GenAI tools work and just what they are doing in your environment.

Agentic AI refers to the family of technology that allows agents—generally using GenAI[1]—to act autonomously to achieve goals with limited human interaction or supervision.

These agents may interact with existing resources (think services already in your environment that expose APIs) using the emerging Model Context Protocol (MCP)[2] standard or through simple resource discovery (e.g., inspecting documentation or code in the environment). Agents may also discover and facilitate communications with each other using the open standard Agent2Agent protocol[3] or through discovery and training.

The **generative AI** at the heart of agents is an extremely powerful technology that relies on its training data (corpus) to create a learned probability. That distribution, when sampled in response to inputs, generates high-probability outputs.

> **Traditional code** asserts that **2+2=4** by using a deterministic ADD operation.
>
> **GenAI** asserts that **2+2=4** because **4** is the most probable token to follow **2+2=.**
>
> **Traditional computing does the math.**
> **GenAI tells you what people usually say.**

Errors in the training corpus and stochastic variations in sampling can result in different—and occasionally incorrect—outputs for the same inputs.

Variations can be introduced because choosing the highest probability token every time doesn't "feel natural." To mimic human thinking, GenAI system engineers have learned to flatten the probability distribution to allow the selection of less likely tokens.

The net effect of this flattening is that GenAI systems will often produce different outputs for the same inputs, trying any of several different paths even when trying to achieve the same goal. Conversely, traditional code does exactly what it's been instructed to, using the same deterministic operations in exactly the same way every time.

Ultimately, it's still just code. But because GenAI is trained on large data corpora recorded from human interactions such as chat transcripts, published documents, and code, it is effective at emulating human communications by producing outputs similar to what humans might do (hence the moniker "artificial intelligence"). However, despite the "sensation" of human will and personality, GenAI tools are still just code, executing instructions against data according to their algorithms.

The agent's algorithms use the probability distribution and current inputs to generate a *probably correct* output (invoking resources, other agents, or interactive users) to arrive at the objective—and do it a little differently every time. In this regard, agents look a lot like human users, varying their approach in the face of current objectives. But agents explore the tools available to them much more quickly and comprehensively than humans can.

Agents test their boundaries at machine speed—making it more critical than ever to implement identity-first best practices to establish and enforce those boundaries.

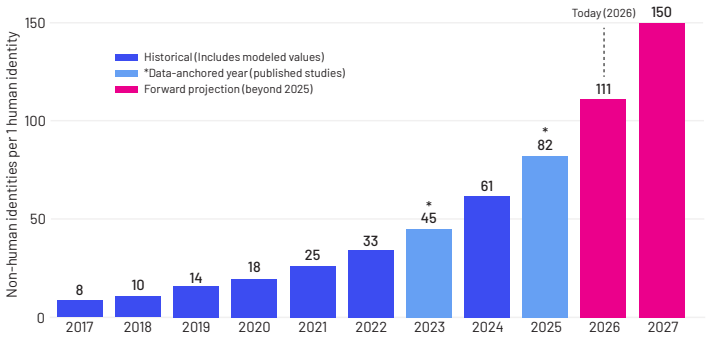
1 Agentic AI vs. Generative AI | IBM
2 What is the Model Context Protocol (MCP)? - Model Context Protocol
3 Agent2Agent

How agents use identity systems to access your stuff

Your applications, services, and other workloads—and now, agents—represent non-human identities (NHIs) in your operational environment. Industry studies show that NHIs already vastly outnumber human identities, and they are increasing exponentially. Microsoft's State of Cloud Permissions Risks[4] and subsequent identity research put the 2018 baseline near 10 NHIs for each human identity (10:1); as of 2026, that ratio was anticipated to be more than 100:1.[5,6]

Global Trend: Non-Human Identities per Human Identity (2017-2027, Illustrative)

Anchored on 2023 and 2025 industry studies; other years are modeled or projected



Modeled global trend based on data from Microsoft's State of Cloud Permissions Risks[4], the Open Worldwide Application Security Project (OWASP)[5], CyberArk's 2025 Identity Security Landscape Report[6], and related industry studies.

Every NHI can potentially be consumed by agents. And agents are themselves consumable—by humans, agents, and other NHIs.

Just like users, applications, and services, agents interact with resources using the authentication mechanism designated by the resource they seek to consume. If a resource responds to an unauthenticated call on a magic port, the agent needn't identify itself at all. If a resource is using secrets (like a certificate or a key), the agent will need access to that secret to access the resource.

If the resource provides access via OAuth, or if MCP is being used and requires authorization, then the agent will require an OAuth access token. MCP specifically adds the Proof Key for Code Exchange (PKCE)[7] to harden security by treating the agent as a public client, binding the authorization to the agent so it is useless elsewhere.

The majority of agents follow one of two authorization patterns: user, or app.

- **If the agent is acting as a user** (e.g., interactive assistants), the authentication flow may require user interaction to get the initial token, typically using OpenID Connect (OIDC).[8] In multi-tiered architecture (e.g., using an Entra ID user token to call the graph), the agent might use the `On-behalf-of` token exchange pattern[9] to call subsequent services with the user's authorization.
- **If acting as an application with its own identity** (e.g., backend infrastructure agents), the agent will authenticate as a service principal, service account, or managed service identity, depending on the platform. In these cases, the agent has its own credentials (sometimes managed by the platform).

Not all resources require authorization, so in some cases, MCP will accept simple API keys or other secrets to access a resource.

In all these cases, risks arise when secrets leak outside the intended security boundary—whether through user or agent social engineering or discovery of an `.ini` file with an embedded secret.

The bottom line is that in most cases agents use identity, represented by accounts, and proven by secrets (either via the user's credentials or their own). Those secrets must be defended lest the rights granted to the entity leak to an attacker.

4 2023 State of Cloud Permissions Risks | Microsoft Community Hub
5 OWASP NHI Top 10 Risks for 2025 Explained by GitGuardian
6 2025 Identity Security Landscape Report | CyberArk
7 RFC 7636 - Proof Key for Code Exchange by OAuth Public Clients
8 Final: OpenID Connect Core 1.0 incorporating errata set 2
9 RFC 8693 - OAuth 2.0 Token Exchange

This makes the need for identity infrastructure greater than ever. Not just because all these new agents will rely on identity systems to prove their own identities (authentication) but also because those identities (and the secrets used to authenticate them) must be protected. Strong identity protection—attack path mapping and detection of attacks and behavioral anomalies indicating compromised accounts—is more critical than ever.

To perform their assigned jobs, agents need to be able to authenticate.

This is an *operational need*.

Agents use identities that must be defended against compromise.

This is a *security need*.

Agents have rights too!

Agents will work within the boundaries available to them. These permissions define what resources an agent can access and what they can do with those resources.

Two primary mechanisms enable access control—whether for humans, agents, or other workloads.

The intersection of these mechanisms defines the permissions available to the agent: first get a token to the resource, then see what the resource will allow.

- **At the resource**. Think in terms of embedded RBAC controls or content restrictions in a CRM or document sharing service.
- **Through the identity system.** Your identity fabric decides whether to grant the agent a token at all (and what scopes to include in the token).

Excessive permissions, excessive risk

Microsoft's *State of Cloud Permissions Risks* report[10] reveals users and apps in most identity systems are substantially over-permissioned. The problem is especially acute for workload identities.

- Only **~1%** of granted permissions are used in daily functions.
- **99%** of accounts are over-permissioned; **half** of these are classified as **high risk**.
- **80%** of workload identities are unused but retain system access—**effectively 100% over-permissioned.**
- **Less than 5%** of permissions granted to workload identities are ever used.

Discovering these over-permissioned identities is a huge windfall for attackers. When malicious actors control an identity's credentials, they can do whatever the identity has permission to do—and the broader the permissions, the better.

Why does the problem of over-permissioning exist?

Because dialing in the *right access* at development or deployment time—especially under deadline pressure—is a lot harder than just requesting *all access* while thinking, *"I'll dial that back soon."*

It's like having that extra piece of cake—and promising yourself you'll exercise it off later. Eating the cake is a lot easier than doing the exercise.

10 2023 State of Cloud Permissions Risks | Microsoft Community Hub

The business pressure that causes over-permissioning in traditional application development propagates to agent development, and low-code/no-code and agentic development brings in a large contingent of new "developers" (including information workers with their new AI subscription) building new tools without the experience, code pipeline support, or predictable code execution that supports traditional application development.

> Whether interacting with legacy applications or new, low/no code or agentic ones, benign agents will exploit excess permissions in unexpected and often harmful ways—while **attacker-controlled malicious agents will enjoy a feeding frenzy in your systems.**

Because identity systems play a crucial role in asserting the rights of agents and the users they act on behalf of, strict identity governance is essential.

- **Rigorously apply** least-privilege, just-enough, just-in-time access—with approvals and reviews—to agents, the resources they consume, and users.
- **Deeply understand** which permissions can grant attacker access to Tier 0 identity systems—directly or through unexpected pathways like nested groups or integrated HR systems.
- **Rapidly detect** permission anomalies at the user, application, or agent scope so you can quickly understand whether rights are being abused by new agentic code paths—or because a malicious actor is controlling the identity.

The care and feeding of agents in your identity fabric

Agents, like all identities, exist in the context of authentication (who they are) and authorization (what they can do). As such, they are highly dependent on the interconnected mix of on-premises and cloud identity providers (IdPs) and authentication protocols that must work together and weave the identity fabric.

If components of that identity fabric are unavailable, agents can't authenticate to gain access to protected resources (including other agents), validate incoming requests from other humans or agents, or be regulated based on their identity by identity-centric security and access control systems like Conditional Access.

> In all mainstream agentic infrastructure, the loss of access to identity infrastructure eliminates agents' ability to operate.

If you depend on agents, you must ensure their operational resilience—and the resilience of the identity fabric on which they operate.

Furthermore, the tools you use to provide this resilience (e.g., tools to automate recovery of compromised identity systems) can't use the same agentic infrastructure deployment that supports the business. Why? Because if your identity systems are offline, you can't access those tools.

Recap

Rogue agents can increase threats against identity infrastructure, while mission-critical agents depend on the resilience of the identity fabric—which can't be recovered by the agents that depend on that fabric.

The Vulnerabilities of Youth

There's an adage that says, "Only time proves security."

There are thousands of examples of companies declaring their systems "perfectly secure" only to be proven wrong when malicious actors discover a new vulnerability or a new attack technique. On any given day, a scan of the tech press will show you a range of containment errors, logging errors, data protection errors, and a variety of new attacks even against systems that have been battle tested and hardened for decades.

But generative and agentic AI are new and rapidly evolving technologies. Many companies are racing to develop or leverage these technologies to build even newer systems – and all this novelty in AI and the technologies developed with it provides fertile ground for exploitation.

It's important to account for this as you prepare generative and agentic deployments. Anchor your initiatives in deep understanding and control of the access granted to agents, users, and workloads.

Social disease

Social engineering is a vulnerability class worth our explicit attention. We primarily think about social engineering of *users by agents* through techniques such as spear-phishing and convincing deep-fake videos, emails, or voice synthesis. But we must also address the social engineering of *agents by users*—or even other agents.

A recent example of this is the early 2026 cyberattack[11] in which Anthropic's Claude Code was weaponized by a single actor to attack 10 Mexican government agencies and a financial institution, exposing approximately 195 million identities. Claude Code initially refused requests that violated its safety guidance.

But the attacker sent more than 1000 prompts framing their activity as legitimate bug-bounty or testing to jailbreak Claude. Claude then participated with the attacker to identify vulnerabilities, generate exploits, plan, and automate the attacks.

Benign agents participating in malicious activity to achieve their objectives is a potential failure mode we should also consider. Somewhere in the training corpus are examples of deception, social engineering, and phishing. So, agents might not just request access to resources to complete their tasks but also tell convincing-but-false stories to human approvers to gain that access.

Once again, it falls to identity system admins to carefully protect agent credentials, strictly and deterministically regulate the permissions an agent has, and detect and mitigate anomalies in behavior.

Social engineering can be applied to agents, just as it can be to users. Admins need to:

- **Apply a least-privilege model** to AI agent access. Actively root out over-permissioning.
- **Leverage User and Entity Behavior Analytics (UEBA).** Look for out-of-norm agent access patterns.

11 Hackers Weaponize Claude Code in Mexican Government Cyberattack - SecurityWeek

The power—and danger—of non-deterministic problem solving

The non-deterministic nature of GenAI is part of what makes this technology powerful—but also potentially dangerous in an environment that's not prepared for its deployment.

A deterministic application will repeat itself, solving a problem with the coded approach exactly the same way every time. In contrast, an agent may discover resources "off the beaten track" and try approaches that aren't typically used each time it tries to achieve its goal.

Imagine the vast collection of resources in your organization as a field.

Valid resources form a path to reach objectives. Other resources—like sensitive docs that are only available because of over-permissioning—are security landmines.

- Well-written **applications** specifically follow the stones in the path to reach the objective.
- **Users** see the path based on resources they've been told they have and generally stick to it because it's easier and because it's the right thing to do.
- But **agents** will wander all over the field, tapping into resources that they have access to without regard to (or in fairness, knowledge of) the fact that access is only a byproduct of poor governance.

> The security debt you've been carrying may have been tolerable because code, regardless of its permissions, just follows its instructions and ignores permission it doesn't need. Humans don't want extra work—and have a sense of ethics—and so generally don't use permissions they don't need. But agents are neither strictly following instructions nor ethical—and will exploit every resource in their grasp.

The end of "security through obscurity"

There was a time when, in some organizations with extensive legacy identity systems, controlling access meant simply hoping that users or applications wouldn't find the resources they shouldn't.

This "security through obscurity" policy was never a good idea, but the problem became more acute with the advent of search indexing for resources. The core difference is:

- **Indexed search** uses string matching—now enhanced with ranking, stemming, synonyms, and machine learning classification.
- **Generative search** tools assess user intention based on semantic analysis and synthesis.

In indexed search, I might look for docs including the key phrase **"2026 financial forecast;"** if a doc instead says "**recent money outlook,"** I might not see it.

With generative search, the system will parse my request by understanding that I mean **"find me recent financial docs,"** then generate a sequence of search returns that are highly likely to match my intent, coalesce the results, and provide me with a summary.

I recently discussed this issue in a webinar.[12]

Here's an example.

You belong to a group that has access to files you shouldn't. This access is largely unexploited because you don't know you have access, nor are you interested in looking. But with generative search, you are more likely to consume data from such resources. Furthermore, if you are using GenAI to produce documents, you may inadvertently include that sensitive information in a document you generate and publish.

12 Top AI Attacks and How ITDR Can Prevent Them

Here's some nightmare fuel.

An over-privileged user account with access to red team data is compromised. The attacker asks a generative search tool to:

- Summarize all current and active vulnerabilities in the environment.
- List detections that might catch exploits of those vulnerabilities.
- Enumerate all known credentials to gain privileged access into identity systems.

Recon has never been easier—and phishing-resistant credentials have never been more important!

The ability for AI to exploit generative search puts a strong mandate on understanding and mitigating the pathways into your core systems. Where permissions exist, generative tools *will* discover and exploit them.

As agentic AI enters enterprise environments with security debt and begins to optimize for outcomes, excess permissions *will* be discovered and exploited. That's certain even in the case of a benign user using a benign agent.

It gets worse when the agent, the operator, or both are malicious.

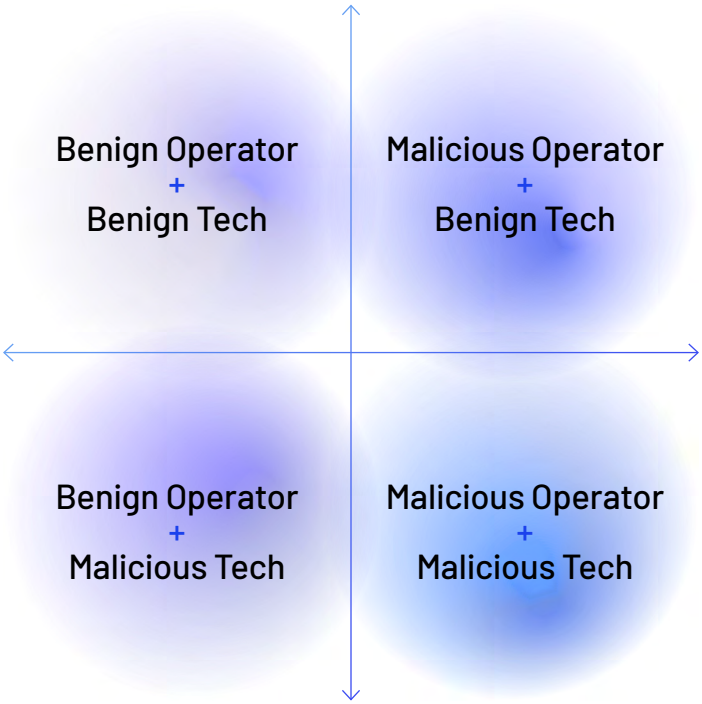
The Agentic Threat Matrix

An agentic threat matrix encompasses the fact that either the operator or the provider of the technology can be malicious or benign. Importantly, even if neither is malicious, these powerful new tools still carry intrinsic risks.

In this matrix, the terms *benign* and *malicious* signify intentions only. For example, a user may be well-intentioned but untrained, and for our purposes is benign.

We explore the many vulnerabilities of well-intentioned systems and users because, of course, perfect does not exist. As defenders, we must not fall into the "I can set it up perfectly and forget about it" trap. We must prepare, drill, and be ready to execute at machine speed.

Let's explore the matrix in depth, examining risks, response, and the role of identity systems.



Benign Operator + Benign Tech

Even when benign operators use benign technology, multiple considerations are in play.

- Is the user over-provisioned with access to resources the agents can leverage?
- Are the policies around inclusion in mail or other external communications being violated?
- Does the underlying platform respect data protection policies (e.g., personally identifiable information, confidentiality)?
- Is the agentic infrastructure leaking data through logs, tenant boundary violations, or other implementation or operational flaws?
- Is the infrastructure that provides the tools compromised?
- Is GenAI-returned data corrupted by results flattening or hallucinations; and are agents propagating that data into other agents or systems?

Tier 0 identity systems contain secrets in documents, email, and other systems that might be visible to agents. A benign user who isn't carefully checking generated content might propagate data useful to attackers outside of expected security boundaries.

As we have seen from OpenClaw deployments,[13] powerful new tools that appeal to users can lead to broad shadow IT deployments. In many cases, neither the operator nor the framework is malicious, but these deployments shatter most of the governance and privilege management intentions of the organization.

Benign Operator + Malicious Tech

AI tools may not be acting with your organization's best outcomes as the goal. They can be compromised in the supply chain, via prompt injection, or deliberately introduced via consent phishing or viral adoption.

Risks include:

- Rapid recon of vulnerabilities to use in building an attack campaign
- Data exfiltration/espionage
- Propagation and placement of persistence mechanisms
- Destruction or encryption of data
- Malicious interactions with or "tricking" of other agents
- AI-assisted consent phishing, credential phishing, and permission phishing
- Automated social engineering attacks

Malicious Operator + Benign Tech

The Claude-facilitated attack on Mexican government agencies provides an example of good tools manipulated to perform attacks, as does our "nightmare fuel" example of an attacker summarizing all known entry points into your organization—with credentials—in seconds.

Other risks include:

- Intentional pollution of training data
- Prompt injections via configuration of local tools
- Data obfuscation
- Use of valid tools for scale abuse (think expense reports)

13 OpenClaw gives users yet another reason to be freaked out about security - Ars Technica

Malicious Operator + Malicious Tech

The operator who introduces code into the environment with the intention to do harm is what we most often think about for at-scale phishing campaigns, infrastructure destruction, or data exfiltration.

We may add:

- Rapid, chained discovery of credentials and permissions
- Rapid discovery of over-permissioned or sensitive data
- Manipulation of data to escape DRM containment

How to respond: Apply foundational principles

To address these types of risks—either benign or malicious—you must gain a deep understanding of how your organization balances risk, convenience, and security. Minimize the assets that are exposed to generative and agentic AI tools as well as the users and applications they interact with.

- Strictly define and enforce permissions.
- Lean into the principles of least privilege and Zero Trust.
- Ensure high fidelity logins.
- Prohibit unknown code access to resources.
- Rapidly detect access and permission anomalies, initiating rollbacks when appropriate.
- Enable policy enforcement by monitoring your identity infrastructure, leveraging tools that provide centralized visibility into who is accessing what.

These foundational best practices—least privileged access, identity protection, and anomaly detection with action interventions—remain essential, regardless of how GenAI is deployed in your environment.

This list leads us to another consideration:

Can AI agents also be our allies in cyber defense?

Won't Agentic "Solve" Resilience for Me?

On February 20, 2026, Anthropic launched Claude Code Security,[14] a tool that scans code bases for security flaws and recommends patches. This is an extremely good use case and has the potential to meaningfully address security debt in legacy code bases.

While the tool was in preview—and although it doesn't directly address identity or use cases like configuration hardening, attack detection, or forensics—markets responded. CrowdStrike stock lost 11% and Okta lost 6% that day; several other cyber stocks were also affected. Markets were sensitive to new approaches disrupting traditional security mechanisms.[15]

That early industry response raises a compelling question: How should we think about agentic AI and its ability to solve for identity resilience?

GenAI and agentic technology will positively impact identity resilience but will likely penetrate the space more slowly than you might expect—and probably will never fully replace identity resilience technology.

Here's why.

- **As AI agents propagate, the identity fabric becomes much more important.** Agentic deployment relies on deterministic policy and governance controls—*all of which are more important in the face of agentic.*
- **Agentic technology relies heavily on the identity fabric.** Agents rely on the identity fabric for authentication and access. While agentic approaches to identity resilience are possible, they can't depend on the same fabric that they are trying to protect at recovery time. *If the identity fabric fails, most mainstream agentic AI fails with it.*
- **The identity fabric is the bedrock of everything in the environment.** Including AI agents. Your identity infrastructure is the last place you want agents acting without oversight. Automation in this environment must provide absolutely predictable outcomes. *Putting critical functions under the control of rapidly emerging probabilistic technology with rapidly emerging security issues is ill advised—to put it mildly.*
- **Critical IAM functions must leverage proven solutions.** Even when agentic has matured enough to be trusted with core identity resilience tasks, ironclad and centralized control of configuration, policy, security detections, and recovery will still be critical. *Core functionality will still rely on deterministic, highly vetted, and trusted subsystems.*

Given these constraints, let's make some predictions regarding the current and future role of GenAI in identity resilience—before, during, and after cyber incidents.

14 Anthropic Launches Claude Code Security for AI-Powered Vulnerability Scanning
15 AI Rattles Cybersecurity Markets: What Anthropic's Code Security Actually Does

Before, During, and After: Addressing Identity Incidents with Agentic AI

Identity resilience requires **preparation before** attacks, **disruption during** attacks, and **recovery after** attacks. Let's examine how agentic AI plays in each of these interconnected phases.

Before the attack

Attacks are ongoing. That means you need continuous environmental assessment and hardening to stay ahead of the barrage.

As important as it is to continuously assess transitive attack paths, remove unnecessary access, and address vulnerabilities, these activities become even more urgent in the face of weaponized generative and agentic AI tools. With AI in the equation, the APR attached to that security debt you're carrying is about to go way up.

It is essential to address excess permissions, unused applications, and weak credentials today, before the attackers with agentic exploit them tomorrow.

How agentic AI can help

AI agents can:

- **Assess your environment**. Agents can quickly triangulate known risks against your current configuration, prevalent attacks, and recommended mitigations.
- **Rapidly address hard-to-clear workload security debt**. Tools like Claude Code Security have the potential for solutions such as converting apps that host user passwords to OAuth or passkeys at scale—or converting apps from service accounts to getting keys from an HSM.
- **Simulate the impact of attacks**. Agents can help defenders prioritize interventions for high-impact events, reducing the downside of mitigations and garnering support for greater automation of remediation measures.

Caution: Agents can also help your attacker do the same assessments—and understand what's unmitigated. You're in a race!

Remember: You still need to ensure that these recommendations align with your environment—especially the parts of your environment not exposed to agentic AI for analysis.

You will also need to convince your sponsors and peers that reducing security debt is a priority. AI agents may be able to help you wordsmith the communications, but quantifying the risks to your organization will remain a very human task.

During the attack

Microsoft reports that in just the first half of 2025, the volume of recorded identity-based attacks rose by 32%.[16] At this volume, identity attacks aren't exceptional event—they are background noise.

Continuously detecting and mitigating these attacks is crucial. You must address them on multiple fronts by using passkeys, just enough/just in time enforcement, request blocking, request rollback, and request quarantining.

You should integrate with your Security Operations Center (SOC) for exceptional cases. But constant, daily handling demands automation—especially given the fallibility of humans and the staffing crisis for security responders.

It also requires shifting your security burden from a small security team to the whole organization, ensuring global application and enforcement of foundational security practices.[17] It's better for a user to have a false-positive that requires them to re-authenticate with their passkey than it is to have a false-negative that compromises your identity infrastructure.

Caution: Used in these ways, agentic becomes part of your Tier 0 defenses and hence is subject to the same attacker attentions as other assets. From an attacker's perspective, a compromised agent is as good as or better than a compromised user.

Remember: Agents can be manipulated by social engineering. For Tier 0 identity systems, we recommend that in addition to agentic AI assistance, you use human"dual key" approval—that is, requiring a human expert to make the call to allow exceptions.

How agentic AI can help

This is one of the most exciting areas for the development of AI-driven defenses and will likely become much stronger in the next few years. AI agents can:

- **Enable rapid triage**. You can create agents that quickly detect use anomalies and abnormal behavior.
- **Empower threat hunting and investigations**. Agentic AI can work alongside traditional supervised and unsupervised classifiers to automate attack detection, reduce attacker dwell time, and help teams intercept and contain attacks.

16 Microsoft Digital Defense Report 2025 | Microsoft
17 ISMG interviews Alex Weinert at RSA Conference 2025

After the attack

As much as we prepare for and mitigate attacks, identity resilience requires that we are prepared for the day when attackers find their way through. Whether the goal is ransomware or espionage, compromise of identity systems is devastating and we must be prepared to recover in a worst case, lights-out scenario.

This level of resilience requires not only having a crisis response plan but also practicing it and ensuring we have the resources needed to execute it—such as immutable backups, out-of-band communications, cleanrooms, data to support identity forensics, and an isolated, trusted environment to work in when identity systems are down.

Caution: AI agents can help to a limited degree in executing the crisis response plan during a real crisis. Again: the systems we depend on for recovery can't be dependent on our compromised identity infrastructure.

Remember: Agents can help with calling the right APIs but when identity is down, they won't provide the functionality behind the APIs that back up, recover, assess, and synchronize identity systems to get themselves and other agents back online.

How agentic AI can help

Agentic can help us:

- **Develop a cyber crisis response plan.** Teams can use GenAI queries and AI agents to create a crisis response and management playbook that goes beyond checking compliance boxes and empowers teams to make effective, confident decisions during a real crisis.[18]
- **Practice the plan.** Your teams can leverage agents to developplausible,complex,high-stakesattacksimulation scenarios, then help you facilitate tabletop exercises that test not only your plan's viability but also your team's readiness.[19]
- **Improve the plan.** After a tabletop exercise, AI-driven analysis of the simulation and response can reveal gaps and provide recommendations for adapting and revising the plan over time.

An AI agent can invoke the cleanroom or recommend a backup, but when the chips are down, the core execution of recovery relies on simple components operating in heavily vetted modalities.

18 Rethinking Cyber Crisis Management: Why Plans Fail
19 Facilitating Cyber Crisis Tabletops: Front-Line Leadership Insights

Planning to Fail with Agentic

GenAI is here, and new AI agents will rapidly proliferate through our environments, whether through intentional adoption or shadow deployments. Securing identity infrastructure for the deployment of agentic technologies depends heavily on known best practices.

It's easy to say that all users should use passkeys, applications should be strictly constrained to their correct execution environment, all permissions should adhere to least privilege principles, and all known vulnerabilities in our identity infrastructure must be mitigated.

But virtually no one can assert that all those things are true in our organizations.

Whether by luck or grace or lack of attacker resources, we have to a greater or lesser extent gotten away with having those gaps. We have been living in a "grace period."

Agentic AI ends the grace period.

Wherever they sit on the threat matrix, agents will discover credentials, exploit excess permissions, and propagate attacks at machine speed. We must do all we can to prepare. We must also prepare to fail. We are all aware of the increasing velocity and effectiveness of attacks.

Achieving cyber resilience has never been more important.

Where appropriate, use agents to help identify and address gaps in your security posture, detect and respond to attacks, and prepare for recovery in the case that attacks break through. And do so with your eyes open to the issues inherent in this developing technology. The efficiencies and benefits of agentic AI are real now and will become even more powerful as the technology matures.

Simultaneously, ensure you are prepared to execute in a lights-out environment in which neither traditional human communication tools nor agents that rely on your compromised identity infrastructure can function.

Whether your tools are operated by humans or humans assisted by agents, an environment that enables collaboration, effective decision making, and rapid response—backed by vetted tools for preparation, detection, and recovery—will always be the core of your strategy.

Plan for identity—and GenAI—to fail, so you can recover with confidence.

About Semperis

Semperis is the identity-driven cyber resilience and crisis management company trusted by the world's largest enterprises and government agencies to protect critical identity systems. Purpose-built for multi-cloud and hybrid identity environments—including Active Directory, Entra ID, Okta, and Ping Identity—Semperis helps organizations prevent, detect, respond to, and recover from identity-based cyberattacks.

Modern cyberattacks are won or lost at the identity layer, where failures now escalate into full-scale business crises. Semperis' AI-powered platform unifies identity lifecycle defense and crisis management—hardening identity infrastructure, detecting and containing active threats, enabling rapid, trusted recovery, and supporting secure, out-of-band coordination when core systems are disrupted—all reinforced by a world-class identity forensics and incident response team.

As part of its mission to help organizations achieve true cyber resilience, Semperis supports the broader cyber community through the award-winning Hybrid Identity Protection (HIP) Conference and Podcast and free identity security tools including Purple Knight and Forest Druid. More than 1,200 organizations—including 25% of the 100 largest U.S. companies—rely on Semperis. The company is privately held, headquartered in Hoboken, New Jersey, and serves customers in more than 40 countries.

About the Author

As Chief Product Officer for Semperis, Alex Weinert drives Semperis' overall product vision and leads strategic innovations that empower organizations to strengthen their hybrid identity security and organizational resilience. He draws on deep experience, including his time as Vice President of Identity Security at Microsoft, where he led teams that protected billions of consumer and enterprise users from unauthorized access, account takeover, and abuse. You can find Alex at events and engagements around the world, connecting, educating, and supporting cyber leaders and defenders in their efforts to build a more secure digital future.

+1-703-918-4884 | info@semperis.com | www.semperis.com

5 Marine View Plaza, Suite 102, Hoboken, NJ 07030