SANS | Research Program

2026 EDITION

Zero Trust and AI in Cloud Environments:

Securing Autonomous Agents and Modern Workloads

Written by:
• Dr. Anton Chuvakin
• Dave Shackleford

In partnership with:

Google

August 2026

About the Authors

Dr. Anton Chuvakin

Google Cloud Security Advisor at Office of the CISO Until June 2019, Dr. Anton Chuvakin was a Research VP and Distinguished Analyst at Gartner for the Technical Professionals (GTP) Security and Risk Management Strategies (SRMS) team. At Gartner, he covered a broad range of security operations and detection and response topics, and is credited with inventing the term "EDR." He is a recognized security expert in the field of SIEM, log management, and PCI DSS compliance. He is an author of the books "Security Warrior," "PCI Compliance," and "Logging and Log Management" and a contributor to "Know Your Enemy II," "Information Security Management Handbook," and others. Anton has published dozens of papers on log management, SIEM, correlation, security data analysis, PCI DSS, and honeypots.

Dave Shackleford
Senior Instructor at SANS Institute

Dave Shackleford is the owner and principal consultant of Voodoo Security. He has consulted with hundreds of organizations in the areas of security, regulatory compliance, and network architecture and engineering, and is a VMware vExpert with extensive experience designing and configuring secure virtualized infrastructures. Dave is a SANS Analyst, serves on the board of directors at the SANS Technology Institute, and helps lead the Atlanta chapter of the Cloud Security Alliance. Dave teaches the following courses at SANS: SEC501: Applied Cyber Defense; and SEC504: Hacker Tools, Techniques, and Incident Handling.

Introduction: The Convergence of AI and Cloud Security

Enterprise cloud environments are entering a new phase of evolution driven by the rapid integration of AI into core business processes. Over the past decade, many organizations have transitioned from traditional on-premises systems to distributed, cloud-native architectures that emphasize scalability, agility, and API-driven interactions. Alongside this shift, security strategies have evolved, with Zero Trust emerging as the dominant model for managing risk in environments where the traditional network perimeter has effectively disappeared.

The introduction of AI—particularly autonomous and semi-autonomous agents—represents the next major inflection point in this evolution. Unlike traditional software systems, which execute deterministic logic based on predefined rules, AI agents operate using probabilistic reasoning and contextual decision-making. They are capable of interpreting intent, generating plans, executing actions across multiple systems, and adapting their behavior based on new information. In doing so, they fundamentally alter how applications interact with data, infrastructure, and users.

This transformation introduces a new set of security challenges. AI agents are not simply applications; they are dynamic actors within the environment. They can initiate actions, access sensitive data, and interact with other systems without direct human intervention. This creates a scenario in which security controls must account not only for access and execution, but also for intent, reasoning, and emergent behavior.

AI agents are not simply applications; they are dynamic actors within the environment.

At the same time, identity has become the central control plane of modern cloud security. Access to resources is governed almost entirely by identity and access management (IAM) systems, making identity the primary target for attackers. The proliferation of AI agents exacerbates this challenge, introducing millions of new identities—many of which are nonhuman, highly privileged, and difficult to monitor.

The convergence of these trends necessitates a new approach to security. Organizations must extend Zero Trust principles to address the unique characteristics of AI-driven systems, ensuring that innovation can proceed without introducing unacceptable levels of risk.

The New Threat Landscape: AI-Powered Threats, Identity, and Agents

The modern threat landscape is evolving rapidly with the latest generation of powerful foundational AI models with advanced cyber capabilities and a shift toward identity-based attacks. In cloud environments, where access to resources is governed by identity rather than network location, attackers have adapted their strategies accordingly. They now have the ability to use AI to discover zero-day attack vectors and can gain legitimate entry to critical systems by compromising credentials, tokens, and access mechanisms.

This approach is highly effective. Once an attacker has obtained valid credentials or tokens, they can often operate within the environment without triggering traditional security controls. This "living off the land" strategy allows them to blend in with normal activity, making detection significantly more difficult.

The introduction of AI agents amplifies this risk. Each agent represents a new identity with its own set of permissions and access pathways. In many cases, these identities are granted broad access in order to perform complex tasks, such as querying databases, interacting with APIs, or managing infrastructure. This creates a large and rapidly expanding attack surface that is difficult to secure using traditional methods.

Each agent represents a new identity with its own set of permissions and access pathways.

In addition to identity-related risks, AI systems introduce entirely new classes of threats. One of the most significant is agent hijacking, in which attackers manipulate the inputs or instructions provided to an AI system in order to alter its behavior. This can occur through prompt injection attacks, malicious data inputs, or compromised upstream systems. Unlike traditional exploits, these attacks target the reasoning processes of the AI itself, making them both novel and difficult to defend against.

Another emerging concern is the proliferation of shadow AI and now shadow AI agents as well. Employees and teams often adopt AI tools without formal approval, integrating them into workflows or uploading sensitive data to external platforms. These activities create unmonitored data flows and introduce significant compliance and security risks. Rogue actions—agents acting not in the interest of its creator due to both attacker compromise and randomness—are coming as well. Organizations often lack visibility into how these tools are being used, what data they are accessing, how outputs are being applied, and what actions they can take on other systems.

Data exposure is also a critical issue. AI systems frequently require access to large datasets in order to function effectively. Without proper controls, this can lead to overexposure of sensitive information, whether through direct access, inference attacks, or unintended outputs. In regulated industries, such as healthcare and financial services, these risks are particularly acute.

Why Zero Trust Must Evolve for AI

Zero Trust has become the foundation of modern cybersecurity strategy, particularly in cloud environments where traditional perimeters no longer exist. The core principle of Zero Trust—never trust, always verify—emphasizes continuous authentication, least-privilege access, and comprehensive monitoring.

However, traditional Zero Trust implementations were designed with certain assumptions in mind. They assume that identities are relatively static, applications behave predictably, and workflows follow well-defined paths. AI systems challenge all of these assumptions.

AI agents are inherently dynamic. They can generate new actions based on context, interact with multiple systems in unpredictable ways, and adapt their behavior over time. This makes it difficult to apply static access controls or predefined policies. Instead, security controls must be capable of evaluating behavior in real time and making context-aware decisions.

Furthermore, AI systems operate at machine speed. They can execute actions far more quickly than human users, which means that any security failure can have immediate and widespread impact. Traditional monitoring and response mechanisms may not be sufficient to detect and mitigate these risks in time.

AI systems operate at machine speed, which means any security failure can have immediate and widespread impact.

Another challenge is the lack of transparency in AI decision-making. Many AI models operate as opaque systems, making it difficult to understand how decisions are being made. This complicates auditing, compliance, and incident response efforts.

To address these challenges, Zero Trust must evolve beyond its traditional focus on access control. It must incorporate mechanisms for validating intent, monitoring behavior, and enforcing policies at runtime. It also must account for the unique characteristics of AI systems, including their autonomy, adaptability, and reliance on external data.

A Zero Trust Framework for AI in the Cloud

The application of Zero Trust principles to AI-driven environments requires a structured framework that addresses both traditional security concerns and the unique challenges introduced by autonomous systems. This framework must be both robust and adaptable, capable of evolving alongside rapidly changing technologies and threat landscapes.

At its core, the framework is built on three foundational principles: human oversight, limited powers, and transparency (see Figure 1). These principles are not new, but their application in the context of AI requires careful consideration and implementation.

Figure 1. Zero Trust Framework for AI

Human Oversight (Governance)

Human oversight remains a critical component of any security architecture, particularly when dealing with systems capable of autonomous decision-making. Although AI agents can perform tasks with remarkable efficiency, they lack the contextual understanding and ethical judgment human operators provide. In practical terms, human oversight involves implementing mechanisms that require human validation for high-risk or irreversible actions. This is often referred to as a human-in-the-loop model. For example, an AI agent tasked with managing cloud infrastructure may be allowed to perform routine maintenance operations autonomously, but any action that could impact production systems or expose sensitive data may require explicit human approval (although this is changing too).

Governance also involves defining clear roles and responsibilities for AI systems. Organizations must establish policies that dictate what actions agents are allowed to perform, under what conditions, and with what level of oversight. These policies should be aligned with business objectives, regulatory requirements, and risk tolerance. Another important aspect of governance is auditability. All actions performed by AI agents should be traceable, with detailed logs that capture inputs, outputs, and decision-making processes. This enables organizations to investigate incidents, demonstrate compliance, and continuously improve their security posture.

Innovation with Control (Limited Powers)

Although AI systems enable new levels of innovation and efficiency, they must operate within clearly defined boundaries. The principle of least privilege is particularly important in this context. AI agents should be granted only the permissions necessary to perform their assigned tasks, and those permissions should be dynamically adjusted based on context and time. Implementing least privilege in AI environments requires a shift from static role-based access controls to more dynamic, context-aware models. For example, an agent may be granted temporary access to a specific dataset for the duration of a task, after which the access is revoked. Similarly, access to APIs and services should be scoped to specific functions, rather than granting broad, unrestricted permissions.

Sandboxing is another critical control. AI agents should operate in isolated environments that limit their ability to interact with other systems. This reduces the risk of lateral movement and prevents compromised agents from affecting the broader environment.

Capability-based security models also can be effective. Instead of granting agents general access to systems, they are given specific capabilities that define what actions they can perform. This approach aligns well with API-driven architectures and allows for fine-grained control over agent behavior.

Transparency (Observability)

Transparency is essential for understanding and managing the behavior of AI systems. Organizations must have visibility into what agents are doing, how they are making decisions, and what data they are accessing. This requires comprehensive logging and monitoring capabilities. Every action performed by an AI agent should be recorded, including the inputs that triggered the action, the decision-making process, and the resulting outputs. These logs should be centralized and integrated with existing security monitoring systems.

Behavioral analytics can be used to identify anomalies and detect potential threats. By establishing baseline patterns for normal agent behavior, organizations can identify deviations that may indicate compromise or misuse. For example, an agent that suddenly begins accessing large volumes of data or interacting with unfamiliar systems may warrant further investigation.

Transparency also extends to the reasoning processes of AI systems. Although it may not always be possible to fully explain complex model decisions, efforts should be made to provide as much insight as possible. This can include logging intermediate steps, providing confidence scores, or using explainable AI techniques.

Runtime Enforcement and Intent Validation

Beyond the foundational principles, modern AI security requires additional layers of control focused on runtime behavior and intent validation. Unlike traditional systems, where access decisions are made at login or session initiation, AI systems require continuous validation throughout execution.

Runtime enforcement involves evaluating actions as they occur and applying policies dynamically. For example, before an AI agent executes an API call, the system can evaluate whether the action aligns with predefined policies, whether the data being accessed is appropriate, and whether the context is consistent with expected behavior.

Intent validation is particularly important in preventing agent hijacking and prompt injection attacks. By analyzing the inputs provided to an AI system, organizations can detect and block malicious instructions before they are executed. This may involve filtering inputs, validating sources, or applying heuristic analysis.

These controls require integration with orchestration layers, API gateways, and service meshes, enabling organizations to enforce policies at critical execution points.

Agent Life Cycle Security and Trust Boundaries

An often-overlooked aspect of securing AI systems is the life cycle management of agents themselves. Unlike traditional applications, which are deployed, updated, and retired through well-defined DevOps processes, AI agents may be created dynamically, modified through prompt changes, or instantiated on demand in response to user queries. This creates significant challenges in maintaining consistent security controls over time.

To address this, organizations must treat AI agents as first-class entities within their security architecture, with clearly defined life cycle stages. These stages typically include creation, deployment, execution, modification, and decommissioning (see Figure 2). At each stage, specific controls should be applied to ensure that agents remain trustworthy and compliant with organizational policies.

AI agents may be created dynamically, modified through prompt changes, or instantiated on demand. This creates significant challenges in maintaining consistent security controls over time.

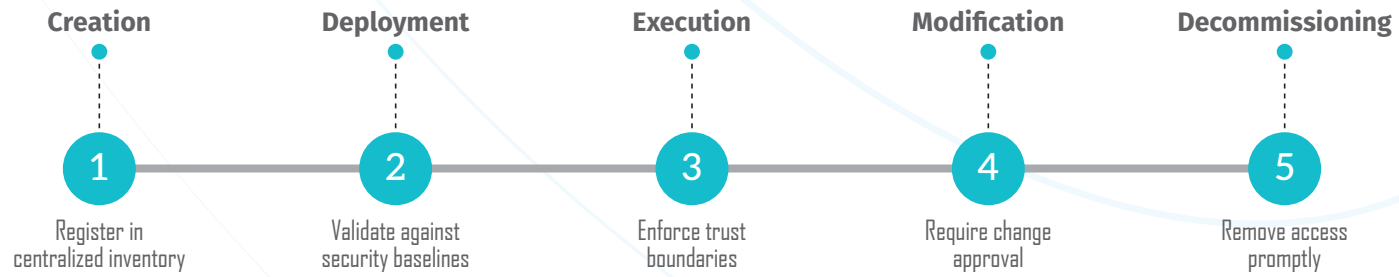

Figure 2. Agent Life Cycle Security

During the creation phase, agents should be registered within a centralized inventory system, with associated metadata that defines their purpose, scope, and permissions. This ensures that all agents are accounted for and can be monitored effectively. In the deployment phase, agents should be validated against security baselines, including checks for appropriate permissions, secure configurations, and adherence to coding or orchestration standards.

Execution represents the most critical phase from a security perspective. During this stage, agents interact with data and systems, making decisions and performing actions. Trust boundaries must be clearly defined to limit the scope of these interactions. For example, an agent operating within a customer support context should not have access to financial systems or infrastructure management APIs. These boundaries should be enforced through a combination of identity controls, network segmentation, and API-level restrictions.

Modification introduces additional risk, as changes to prompts, models, or integration points can alter agent behavior in unpredictable ways. Organizations should implement change management processes that require validation and approval for modifications, particularly those that affect permissions or data access.

Finally, decommissioning is essential to prevent the accumulation of unused or orphaned agents. Just as with user accounts, agents that are no longer needed should be removed promptly, along with any associated credentials or access rights.

By managing the full life cycle of AI agents and enforcing trust boundaries at each stage, organizations can significantly reduce the risk of unauthorized access, unintended behavior, and persistent threats.

Architectural Controls for Securing AI Workloads

Implementing the Zero Trust framework for AI requires a robust set of architectural controls. Cloud platforms provide a range of services that can be leveraged to enforce these controls, particularly in environments built on Google Cloud Platform (see Figure 3).

Identity and access management is a foundational component. GCP IAM enables organizations to define fine-grained permissions for users and service accounts. Features such as IAM Conditions allow for context-aware access control, enabling policies based on attributes such as time, location, and request parameters.[1] Workload Identity Federation provides a mechanism for securely accessing resources without the need for long-lived credentials, reducing the risk of credential leakage.

Secure execution environments are also critical. Google Kubernetes Engine provides a managed platform for deploying containerized workloads, with built-in support for network policies and workload isolation.[2] Confidential VMs[5] offer hardware-based isolation using Trusted Execution Environments, protecting sensitive data and computations from unauthorized access.[3]

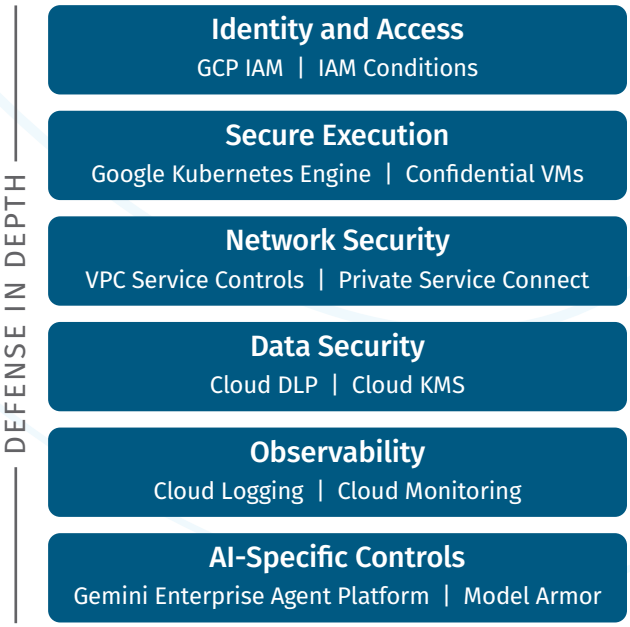

Figure 3. Architectural Controls for Securing AI Workloads on Google Cloud

[1] https://docs.cloud.google.com/iam/docs/conditions-overview
[2] https://cloud.google.com/kubernetes-engine
[3] https://docs.cloud.google.com/confidential-computing/confidential-vm/docs/confidential-vm-overview

Network security controls help prevent data exfiltration and unauthorized access. VPC Service Controls create security perimeters around sensitive resources, limiting access based on network boundaries. Private Service Connect enables private connectivity to Google services, reducing exposure to the public internet.

Data security is another key consideration. Cloud DLP[4] can be used to identify and protect sensitive data, while Cloud KMS provides centralized key management for encryption. These services can be integrated into AI workflows to ensure that data is protected at all stages of processing.

Observability is particularly important in AI pipelines, as issues may arise at any stage. Cloud Logging[5] and Cloud Monitoring can be used to track data flows, model interactions, and system performance. Integrating these logs with security analytics platforms allows organizations to detect anomalies and respond to incidents more effectively.

Finally, AI-specific controls can be implemented using Gemini Enterprise Agent Platform (formerly known as Vertex AI), which provides capabilities for model deployment, monitoring, and governance.[6] Features such as model monitoring and guardrails help ensure that AI systems operate within defined parameters and do not produce harmful or unintended outputs.

Beyond individual workloads, securing AI in cloud environments requires a focus on the broader pipeline through which data, models, and actions flow. In Google Cloud environments, this often involves services such as Gemini Enterprise Agent Platform pipelines,[7] Cloud Functions, Cloud Run, and event-driven architectures built on Pub/Sub. AI pipelines introduce unique risks because they connect multiple components, including data ingestion systems, model training environments, inference endpoints, and downstream applications. Each stage of the pipeline represents a potential attack surface, and a weakness in any component can compromise the entire system.

One critical control is ensuring that data flowing through pipelines is validated and sanitized with services such as Model Armor before it is processed by AI models.[8] This helps prevent prompt injection and data poisoning attacks. For example, data ingested through Pub/Sub topics should be filtered and inspected before being passed to Agent Platform models, reducing the likelihood that malicious inputs will influence model behavior.

Identity controls also must extend across the pipeline. Service accounts used by Cloud Run[9] or Cloud Functions[10] should be tightly scoped, with permissions limited to specific resources. Workload Identity Federation can be used to eliminate the need for long-lived credentials, ensuring that access is granted dynamically and securely.

Finally, orchestration layers should include policy enforcement points that evaluate actions before they are executed. For example, before a Cloud Function triggers an API call based on AI output, a policy engine could validate the request against predefined rules, ensuring that it aligns with organizational policies and does not introduce unintended risk.

Workforce identity was hard. Non-human identity made it harder. Agentic identity makes it something else entirely.

[4] http://cloud.google.com/security/products/dlp
[5] https://cloud.google.com/logging
[6] https://cloud.google.com/products/gemini-enterprise-agent-platform
[7] https://cloud.google.com/products/gemini-enterprise-agent-platform
[8] https://cloud.google.com/security/products/model-armor
[9] https://cloud.google.com/run
[10] https://cloud.google.com/functions

Implementation Roadmap

Implementing a Zero Trust architecture for AI-driven cloud environments is not a single initiative but a multiphase transformation that spans people, processes, and technology. Organizations that approach this effort as a tool deployment exercise are likely to struggle. Success depends on aligning security controls with operational workflows and business objectives. The path toward that alignment begins with three foundational phases, followed by the ongoing work of securing data, establishing observability, and maturing governance (see Figure 4).

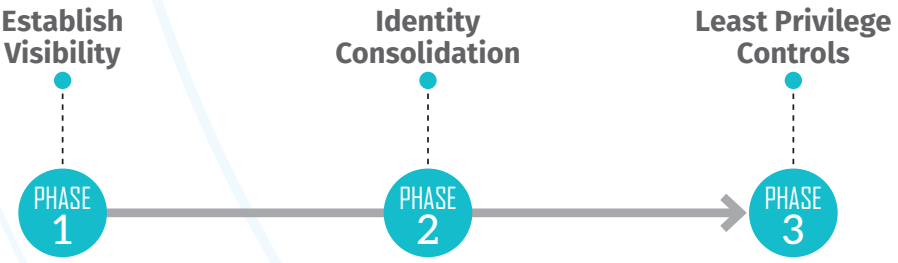

Figure 4. Three Foundational Phases of Alignment

1. **The first phase of implementation involves establishing visibility.** Many organizations do not have a comprehensive understanding of where AI systems are being used, what data they are accessing, or how they are integrated into business processes. A detailed inventory should be conducted to identify all AI agents, associated identities, data flows, and dependencies. This includes both sanctioned and unsanctioned systems, as shadow AI usage often represents a significant blind spot.

2. **The second phase involves identity consolidation and governance.** Once visibility is established, the next step involves centralizing identity management for both human and nonhuman identities, ensuring that all AI agents are registered, authenticated, and subject to consistent policies. Service accounts and API keys should be audited, with unnecessary or overly permissive accounts removed or restricted. Credential rotation policies should be enforced to reduce the risk of long-lived secrets being compromised.

Organizations that approach this effort as a tool deployment exercise are likely to struggle.

3. **The next phase involves implementing least-privilege controls.** This requires a shift from static role-based access models to more dynamic, context-aware approaches. Permissions should be granted based on specific tasks and revoked when no longer needed. This may involve the use of short-lived tokens, just-in-time access mechanisms, and fine-grained API controls. For AI agents, this means defining clear boundaries around what actions they are allowed to perform and ensuring that those boundaries are enforced consistently.

Data security also must be addressed as part of the implementation process. Sensitive data should be classified and labeled, with access controls applied based on classification levels. Data access patterns should be monitored to detect anomalies, such as unusually large queries or access from unexpected locations. In addition, organizations should consider implementing data minimization strategies, ensuring that AI systems only access the data necessary for their intended purpose.

Observability is another critical component of the implementation roadmap. Organizations must deploy logging and monitoring solutions that provide visibility into AI behavior, including inputs, outputs, and decision-making processes. These logs should be integrated into existing security operations workflows, enabling analysts to detect and respond to potential threats. Behavioral analytics can be used to establish baselines and identify deviations that may indicate compromise.

Governance frameworks must be established to define policies and procedures for AI usage. This includes setting guidelines for acceptable use, defining approval processes for high-risk actions, and establishing accountability for AI-driven decisions. Regular audits should be conducted to ensure compliance with these policies and to identify areas for improvement.

Finally, organizations should adopt a continuous improvement approach. The threat landscape is constantly evolving, and security controls must be updated accordingly. This involves regularly reviewing policies, updating configurations, and incorporating lessons learned from incidents and audits. By treating Zero Trust for AI as an ongoing program rather than a one-time project, organizations can maintain resilience in the face of changing risks.

As organizations move from initial deployment to operational maturity, the implementation of Zero Trust for AI must evolve to include deeper integration with existing enterprise processes. One of the most critical aspects of this phase is aligning security controls with development and DevOps workflows. AI systems are often built and deployed rapidly, and security must be embedded into these processes to avoid becoming a bottleneck.

This can be achieved by integrating security checks into CI/CD pipelines. For example, before an AI model or agent is deployed, automated checks can validate its configuration, permissions, and dependencies. These checks should include verification of least-privilege access, validation of data sources, and confirmation that logging and monitoring are enabled. By shifting these controls left, organizations can catch potential issues early and reduce the risk of misconfigurations reaching production environments.

Another important consideration is the role of security operations teams in monitoring and responding to AI-related incidents. Traditional SOC workflows may not be sufficient to handle the complexity of AI systems, particularly when dealing with issues such as prompt injection or anomalous agent behavior. Organizations should invest in training and tooling that enable analysts to understand and investigate AI-specific threats.

This may include developing new playbooks for incident response that address scenarios such as agent hijacking, data leakage through AI outputs, or unauthorized access via nonhuman identities. These playbooks should define clear steps for containment, investigation, and remediation, ensuring that teams can respond effectively to emerging threats.

Organizations that approach this effort as a tool deployment exercise are likely to struggle. Success depends on aligning security controls with operational workflows and business objectives.

Organizations should establish measurable outcomes to evaluate whether Zero Trust controls are effectively reducing AI-related risk. Rather than focusing solely on technical implementation metrics, leaders should track indicators that demonstrate improved control over AI identities, data access, autonomous actions, and incident response. Examples are outlined in Table 1.

Metric	What It Measures
Percentage of AI Agents Under Verified Least-Privilege Policies	Measures how many AI agents, assistants, and automated workflows have access rights limited to only the systems, data, APIs, and tools required for their intended functions.
AI Identity Risk Score	Tracks the number of AI agents with excessive permissions, standing privileges, unmanaged credentials, or access to high-value resources.
Sensitive Data Exposure Rate	Measures the volume of regulated, confidential, or proprietary data accessible to AI systems vs. the volume protected by classification, masking, tokenization, or policy enforcement controls.
Percentage of AI Interactions Subject to Continuous Monitoring	Tracks how much AI activity, including prompts, tool use, API calls, retrieval operations, and autonomous actions, is logged, monitored, and correlated with security analytics platforms.
Autonomous Action Approval Rate	Measures the percentage of AI-initiated actions that require human approval, policy validation, or additional authentication before execution, particularly for privileged or high-impact operations.
Mean Time to Detect (MTTD) and Respond (MTTR) for AI Incidents	Quantifies how quickly organizations identify and contain prompt injection attacks, unauthorized data access, agent misuse, credential abuse, or other AI-related threats.
AI Policy Violation Frequency	Tracks the number of instances where AI systems attempt to access unauthorized resources, invoke prohibited tools, exceed permission boundaries, or violate established governance policies.
Agent-to-Resource Trust Ratio	Measures the number of approved trust relationships between AI agents and enterprise resources, helping identify excessive connectivity and opportunities to reduce attack surface.
Reduction in Standing Privileges for AI Workloads	Evaluates progress toward replacing persistent access with just-in-time, ephemeral, or dynamically authorized access models.
AI Blast Radius Score	Estimates the maximum number of systems, datasets, identities, or business processes that could be impacted if a single AI agent, model, credential, or orchestration platform were compromised.
Percentage of AI Systems Covered by Zero Trust Controls	Tracks adoption of identity verification, device posture validation, microsegmentation, policy enforcement, data protection, and continuous monitoring across the organization's AI ecosystem.
Rate of Blocked High-Risk AI Requests	Measures how often policy engines successfully prevent sensitive data exfiltration, unauthorized model access, excessive tool use, or risky autonomous actions before they occur.

Table 1. Matching Escalation Level to Action Risk and Confidence

Future Considerations

As AI technologies continue to evolve, the security challenges associated with them will become increasingly complex. Organizations must anticipate these changes and prepare for a future in which AI systems are deeply embedded in every aspect of enterprise operations.

One of the most significant trends is the rise of AI-to-AI interactions. In these scenarios, multiple AI agents communicate and collaborate to achieve a common goal. Although this can improve efficiency and scalability, it also introduces new risks. Compromising a single agent could allow an attacker to influence an entire network of systems, creating cascading effects that are difficult to contain. Securing these interactions will require new protocols for authentication, authorization, and trust establishment between agents.

Another emerging trend is the use of AI in security operations. AI-driven tools are increasingly being used to detect threats, analyze data, and automate responses. Although this can enhance security capabilities, it also introduces new attack vectors. Adversaries may attempt to manipulate AI models used in security systems, causing them to misclassify threats or ignore malicious activity. Ensuring the integrity and reliability of these systems will be critical.

Compromising a single agent could allow an attacker to influence an entire network of systems, creating cascading effects that are difficult to contain.

Regulatory and compliance requirements are also likely to evolve in response to the growing use of AI. Governments and industry bodies are beginning to develop frameworks for AI governance, focusing on issues such as transparency, accountability, and data protection. Organizations must stay informed about these developments and ensure that their security practices align with emerging standards.

The concept of digital identity also will continue to evolve. As the number of nonhuman identities grows, organizations will need more sophisticated tools for managing and securing them. This may include identity graphing, behavioral profiling, and advanced analytics to detect anomalies. Identity will remain the central control plane of security, and investments in identity management will be critical.

One of the most significant emerging concerns is the rise of adversarial AI, in which attackers deliberately craft inputs designed to manipulate or deceive AI systems. These attacks may target not only individual models but also entire pipelines, exploiting weaknesses in data validation, model training, or inference processes. Adversarial AI techniques can be used to bypass detection systems, manipulate decision-making processes, or extract sensitive information from models. For example, carefully crafted inputs may cause a model to reveal training data or produce outputs that violate policy constraints. Defending against these attacks requires a combination of robust model design, input validation, and continuous monitoring.

Finally, the concept of autonomous security is likely to gain traction in the coming years. Just as AI is being used to automate business processes, it also can be applied to security operations. AI-driven systems may be able to detect threats, analyze data, and respond to incidents with minimal human intervention. This has the potential to improve efficiency and effectiveness, but it also introduces new risks, particularly if these systems are themselves compromised or manipulated.

The pace of innovation in AI will continue to accelerate. New models, architectures, and use cases will emerge, each with its own set of security implications. Organizations must adopt a flexible and adaptive approach to security, one that can accommodate new technologies without compromising existing controls.

Conclusion

The integration of AI into enterprise cloud environments represents a transformative shift in how organizations operate and innovate. While this shift offers significant benefits, it also introduces new risks that cannot be addressed using traditional security models.

Zero Trust provides a strong foundation for addressing these challenges, but it must be extended to account for the unique characteristics of AI systems. By focusing on governance, least privilege, and observability, organizations can build a comprehensive security framework that supports innovation while protecting critical assets.

Ultimately, the goal is to create an environment in which AI systems can operate safely and effectively, enabling organizations to harness their full potential without exposing themselves to unnecessary risk.

SANS | Research Program

About the SANS Research Program

The SANS Research Program is a key initiative by the SANS Institute and a premier global provider of cybersecurity research and information. SANS Research Program is designed to provide cybersecurity practitioners and leaders with data-driven insights, thought leadership, and solutions that help them better understand and respond to evolving security challenges. All content is authored by SANS instructor experts from around the world who apply their years of experience from hands-on practitioner work in the field, advisory roles, and the classroom to provide education, guidance, and actionable insights that help make the cyber world a safer place.

To learn about sponsorship opportunities for research, content, and in-person or virtual events, email us at **Sponsorships@sans.org** or go to **www.sans.org/sponsorship**.

Free Resources

Access more free educational content from SANS at:

SANS Reading Room
www.sans.org/white-papers

Check out upcoming and on-demand webcasts:

X
x.com/SANSInstitute

LinkedIn
www.linkedin.com/company/sans-institute

YouTube
www.youtube.com/@SANSInstitute