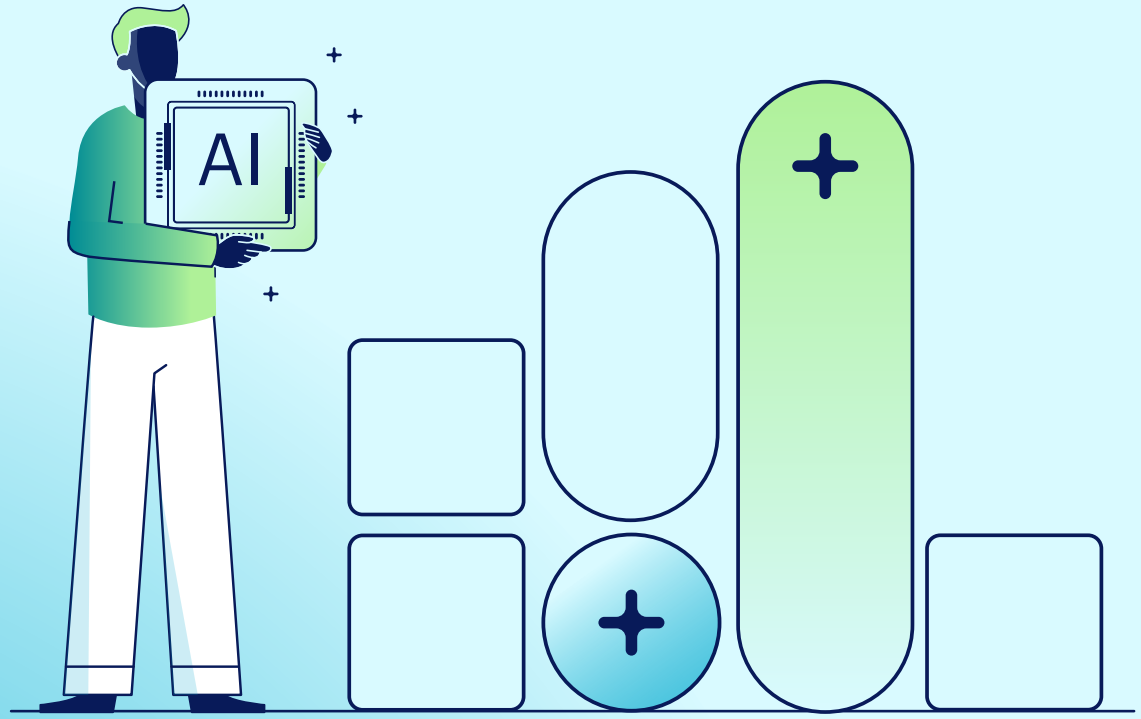
netskope

The AI Security Playbook

A Practical Guide to Securing AI End-to-End, Everywhere

The

AI Security Playbook



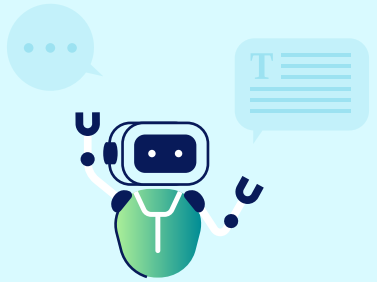


Table of Contents

Introduction

Artificial intelligence (AI) technologies have rapidly established themselves as useful and important tools for many organizations. With new capabilities and use cases emerging all the time, AI is now a core component of most enterprise technology stacks.

AI's rapid ascent has been marked by high levels of investment too. According to analysts at IDC, the worldwide AI IT spending market is estimated to rise to almost $750Bn by 2028 with generative AI specific expenditure being just over $300Bn.[1]

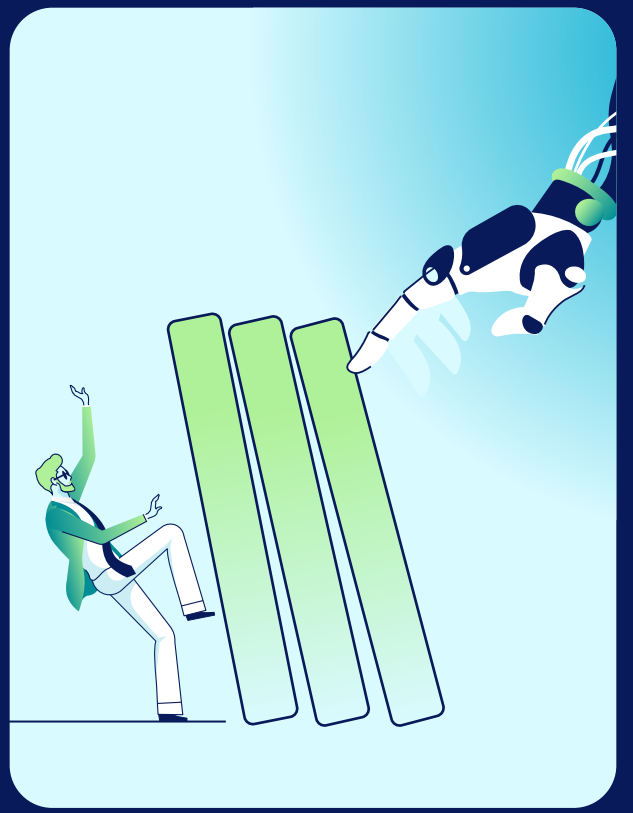
The worldwide AI IT spending market is estimated to rise to almost $750Bn by 2028 with generative AI specific expenditure being just over $300Bn.

For security practitioners, the potential risks of AI applications in their environment are obvious and growing. At the most basic adoption level, data is being shared with third party applications in the cloud. From a security perspective, this raises questions about what data employees are putting into these systems and what controls are in place to manage it. The advance of standard protocols to make data sharing with AI applications even easier—such as Model Context Protocol or MCP—systematize these risks.[2]

Security challenges are likely to intensify further as enterprise AI technology evolves. Agentic AI systems, for example, can operate autonomously to achieve specific objectives or execute defined tasks without requiring constant human intervention. Industry analysts at Gartner forecast that, by 2028, 25% of enterprise breaches will be tied to AI agent abuse.[3]

Given the fast-developing risks facing security professionals, it's no surprise that they're looking for help navigating this new landscape. In this eBook we describe the top security concerns facing organizations today and the solutions that Netskope can provide to help.

Gartner forecasts that, by 2028, 25% of enterprise breaches will be tied to AI agent abuse.



[1] IDC Market Forecast, Worldwide Artificial Intelligence IT Spending Forecast, 2024–2028, Rick Villars et al., October 2024, Doc #US52635424..

[2] Netskope Cloud and Threat Report, 2025 https://www.netskope.com/netskope-threat-labs/cloud-threat-report/cloud-and-threat-report-2025

[3] Gartner's Top Predictions for 2025.

Security Challenges in AI

Top three issues facing security teams today

1 Expanding risk surface

As the use of AI evolves from pure-play generative AI tools (like ChatGPT) to integrated AI capabilities across enterprise apps and privately built AI applications, the attack surface continues to expand. Each stage introduces new risks:

- Public genAI tools introduce risks of inadvertent sensitive data exposure.
- Integrated AI features in existing SaaS apps may open paths for data leakage or manipulation.
- Privately hosted LLMs and custom AI apps introduce new vectors, such as misconfigured access controls or vulnerabilities in data pipelines.
- Connections between AI applications and data sources via new protocols, such as MCP, expand the risk surface for potential data exfiltration.

2 Sensitive data exposure and exfiltration

The most immediate risk in AI adoption is data loss, whether it's accidental or malicious:

- Inadvertent exposure happens when employees input sensitive data (e.g., PII, trade secrets, regulated data) into public models without realizing the consequences.
- Malicious insiders or external attackers may exploit AI tools to exfiltrate data or abuse the model's output channels.
- There's also a training risk: Using improperly curated data in model training can lead to models that leak confidential information.

3 Responsible AI governance

As AI systems scale, they raise critical compliance and ethical concerns that intersect with security:

- AI models can unintentionally encode and propagate biases, leading to regulatory scrutiny and reputational damage.
- Improper handling of employee or customer data used in AI workflows may violate GDPR, HIPAA, or other data privacy laws.
- The autonomous deployment of AI in place of human decision-making, especially in high-stakes areas (e.g., hiring, security, finance), introduces ethical dilemmas and accountability gaps.

AI Security Foundations

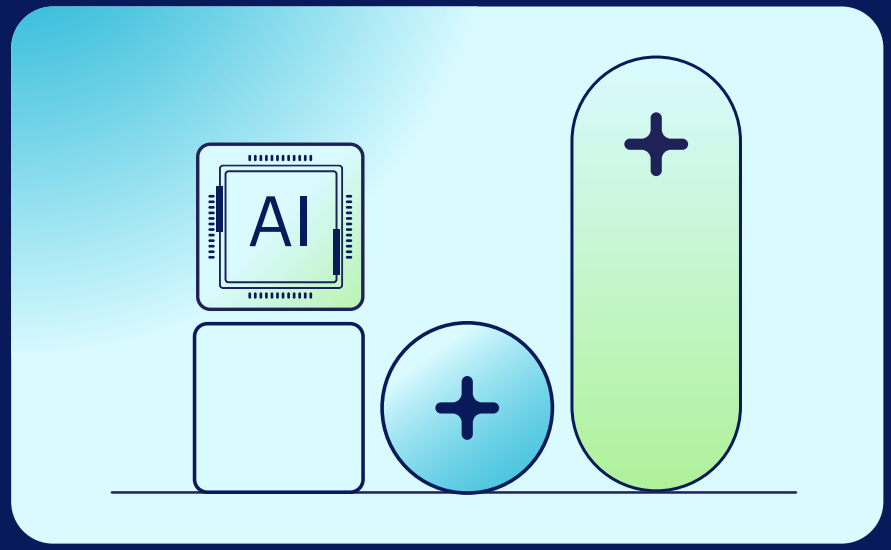
The zero trust imperative

AI security relies on a zero trust approach, much like SaaS security, but with unique challenges that stem from the way AI models process inputs and generate outputs.

Both AI and SaaS security demand strict access controls, continuous monitoring, and robust data protection to mitigate risks. However, while SaaS security primarily focuses on safeguarding applications and user interactions, AI security must also account for the integrity of training data, model access, and potential adversarial manipulation. This makes enforcing context-aware security policies and real-time threat detection essential in preventing data leakage, unauthorized access, and AI model exploitation.

A strong zero trust framework for AI security ensures that every request is verified, every data flow is monitored, and access is granted based on dynamic risk assessments rather than static permissions. This approach requires granular visibility into data movement and adaptive security controls that adjust based on real-time context.

With zero trust principles in place, businesses can securely adopt and scale AI-driven technologies without compromising security or compliance.



Pro Tip

AI security relies on a zero trust approach, much like SaaS security, but with unique challenges that stem from the way AI models process inputs and generate outputs.

Navigating AI Security

Top six challenges and solutions

Challenge #1: Lack of Visibility

As AI tools become embedded in daily workflows, organizations are grappling with a fundamental security challenge: they can't secure what they can't see.

Employees access sanctioned and unsanctioned applications with both corporate and personal credentials, blurring the lines between approved and unapproved use. This uncontrolled sprawl increases the risk of data leakage, IP loss, and compliance violations, especially when sensitive information is entered into unmanaged or shadow AI services.

Most organizations lack the granular visibility needed to differentiate between risky and legitimate AI use. Traditional tools fall short in identifying specific AI model interactions, distinguishing personal from corporate accounts, and providing real-time insights at the user, app, or activity level. Without deep visibility into how and where AI is being used, security teams are left blind to potential exposure points.

How Netskope solves for this

As organizations increasingly adopt AI tools, maintaining visibility and control over their usage becomes crucial. Netskope offers a comprehensive solution to track both managed and unmanaged (shadow) AI applications, providing security teams with the insights they need to ensure proper oversight.

Key capabilities include:

- **Advanced Instance Awareness:** Distinguish between personal and corporate instances of AI applications like ChatGPT, Gemini, and Copilot.
- **AI Dashboard**: Gain deep insights into AI usage trends, top applications, frequency of access, and granular user actions such as logins, posts, uploads, and downloads.
- **User and Entity Behavior Analytics (UEBA):** Detect anomalies and risky behavior using machine learning to identify threats such as data exfiltration, insider risks, and policy violations.
- **Foundational Visibility:** Gain visibility over your AI ecosystem from user-to-app, API and MCP traffic. Netskope unifies visibility from usage, inventory, and data flows.

This holistic visibility enables security teams to act swiftly and mitigate risks tied to AI usage across the enterprise.

Challenge #2: Understanding AI Application Risk

As AI capabilities rapidly evolve, so does the risk landscape. What was once a simple SaaS application can now quietly introduce embedded AI features such as copy generation, smart replies, and AI copilots, without notifying users or security teams. This growing trend makes it increasingly difficult to understand which applications are using AI, how they're using it, and what risks they introduce to the organization.

Security teams need the ability to dynamically assess risk based on how AI features are integrated, whether they retain or train on enterprise data, and how they align with compliance requirements. Without this level of insight, organizations risk exposure to data leaks, intellectual property theft, regulatory violations, and even AI model manipulation. As the AI footprint within SaaS continues to expand, understanding application risk is not just a best practice, it's a necessity for any organization looking to adopt AI safely.

How Netskope solves for this

Netskope tackles the evolving complexity of AI application risk with its Cloud Confidence Index (CCI), which provides real-time, continuously updated insights into more than 85,000 cloud and SaaS applications. With dynamic, AI-aware risk assessments, CCI helps security teams stay ahead of potential risks and ensure compliance.

Key capabilities include:

- **Real-Time, AI-Aware Risk Scoring:** Identify applications with embedded AI capabilities and understand the risks associated with these features.
- **Enterprise Data Handling Insights:** Evaluate how applications manage enterprise data, including retention, model training, and third-party sharing.
- **Compliance Tracking:** Stay aligned with regulatory requirements such as GDPR, SOC 2, and ISO 27001.
- **Secure LLMs and MCP:** Evaluate more than 85,000 SaaS apps, including AI apps and embedded AI features, and public MCP servers, identifying risky attributes, authentication types, and protocol versions.

With CCI, security teams can confidently navigate the complexities of AI application risks and ensure their organization remains secure and compliant.

Challenge #3: AI Model Integrity

As organizations increasingly leverage generative AI tools (both custom-built models and corporate applications like Microsoft Copilot), ensuring the integrity of the data used to train these models becomes a critical concern. These AI systems are often trained on vast datasets that may include sensitive corporate documents, emails, presentations, spreadsheets, and proprietary business information.

If sensitive or proprietary data is inadvertently incorporated into training datasets, it can lead to exposure not only through model outputs but also through adversarial prompts, data leakage, and potential compliance violations. As the adoption of genAI expands across various departments, it becomes increasingly difficult for security teams to control how training data is sourced, validated, and protected.

+ Microsoft Copilot can be trained on the content within a user's Office suite, ranging from Word documents to Excel spreadsheets. If confidential or sensitive data is stored in these locations, and access controls are not properly configured, then there is a potential risk that Copilot could surface sensitive business strategies, financial details, or customer information in its responses.

How Netskope solves for this

Netskope One DSPM (Data Security Posture Management) empowers organizations to monitor and protect sensitive data across cloud environments and data repositories. By detecting and classifying critical data, such as financial records, PII, and intellectual property, Netskope ensures this information is not used to train AI models without proper authorization.

Key capabilities include:

- **Continuous Monitoring of Cloud Environments:** Detect and classify sensitive data in real time, ensuring unauthorized use in AI model training is prevented.
- **Visibility into Data Access and Sharing:** Gain real-time insights into how data is accessed and shared across the cloud, allowing for immediate corrective action when necessary.
- **Compliance and Data Leakage Prevention:** Safeguard sensitive data to ensure compliance, prevent data leakage, and maintain control over intellectual property.
- **Robust Security Posture Management:** Ensure proper data posture and discover, label, and classify your structured and unstructured data.

With Netskope One DSPM, organizations can proactively protect their sensitive data, ensuring AI model training stays secure, compliant, and controlled.

Challenge #4: Threats Targeting AI Systems

Adversaries are evolving their tactics to exploit AI-specific vulnerabilities, using prompt injection, data poisoning, and adversarial inputs designed to skew results or exfiltrate sensitive data. Additionally, AI applications are often integrated with broader business systems, making them a potential entry point for lateral movement, privilege escalation, or data theft.

Whether it's a threat actor trying to manipulate the output of an AI model, extract training data, or exploit weak access controls around AI APIs, the attack surface is expanding rapidly. Compounding this issue is the lack of standard security frameworks for protecting AI systems, leaving many organizations unprepared to defend against novel attack vectors. As AI adoption increases, so does the need for security teams to proactively detect and mitigate threats that specifically target AI environments, before those threats compromise sensitive data, operations, or decision-making processes.

How Netskope solves for this

Netskope tackles the growing threats targeting AI systems with a multi-layered security approach that integrates advanced threat protection, deep visibility, and AI-specific defenses.

Key capabilities include:

- **Unified AI Defense:** Netskope One AI Guardrails mitigates sophisticated attacks—including prompt injection and jailbreak attempts—through deep, real-time analysis of all traffic.
- **Advanced Threat Protection:** Use machine learning, sandboxing, and heuristic analysis to detect and block both known and zero-day threats, including malware hidden in files submitted to AI tools.
- **Red Teaming and Vulnerability Assessments:** Automate adversarial simulations to uncover vulnerabilities, ensuring your private models are secure, compliant, and resilient against advanced threats with Netskope One Red Teaming.
- **Proactive AI Activity Monitoring:** Detect emerging threats and vulnerabilities with real-time monitoring of AI interactions to ensure a comprehensive defense strategy.

By combining these technologies, Netskope provides an integrated solution that helps organizations secure their AI systems from sophisticated cyber threats and evolving attack vectors.

Challenge #5: Data Exposure

One of the most urgent and high-stakes challenges in AI security is the risk of data exposure. As employees across departments adopt AI tools to boost productivity, they may unknowingly upload or share sensitive data such as source code, customer records, financial documents, or proprietary IP with public AI models. Once exposed, this data can be retained, used for model training, or even leaked, depending on the application's privacy policies and data handling practices.

Unlike traditional data sharing channels, AI platforms can act as black boxes, offering little transparency into how data is stored, accessed, or used. Without guardrails in place, organizations face serious risks ranging from regulatory violations and IP theft, to reputational damage and competitive disadvantage.

+ Netskope Threat Labs observed the exposure of source code in nearly 50% of AI-related policy violations. This underscores how easily critical business assets can be compromised through seemingly benign actions, like pasting a snippet of code into an AI chatbot to debug or optimize it.

How Netskope solves for this

Netskope provides comprehensive, context-rich protection for enterprise data, at rest and in motion. By combining real-time risk assessments, inline and API-based controls, and posture checks, Netskope's unified security policies enable precise governance of both user and data interactions across the organization.

Key capabilities include:

- **Advanced Data Loss Prevention (DLP):** Protect sensitive information from exfiltration through AI tools whether users are in the office, at home, or on the go.
- **Granular Control:** Block or limit high-risk actions, such as uploading source code or confidential documents.
- **Real-Time User Coaching:** Educate users on policy violations with visual prompts, helping to reduce repeat offenses.
- **Inspect every request and response:** Identify and block the delivery of patented or copyrighted data in AI responses to proactively defend against IP risks associated with generative model outputs.
- **Secure API Traffic:** Authenticate and centralize traffic management and content inspection between private apps and LLMs.

With these capabilities, Netskope ensures comprehensive, adaptive data protection that scales across an organization's entire AI and cloud environment.

Challenge #6: Governance, Compliance, and Ethical Use

As AI adoption accelerates, organizations face growing pressure to align with emerging governance standards, regulatory requirements, and ethical expectations, especially in highly regulated industries like finance, healthcare, and government. Countries around the world are rapidly introducing AI-specific frameworks and mandates, seen in the EU AI Act, NIST AI Risk Management Framework, and U.S. executive orders on AI safety. These regulations aim to ensure responsible development and deployment of AI systems, mandating transparency, data privacy, explainability, and non-discrimination.

However, meeting these standards is no easy task. Security and compliance teams must understand how AI is used across their environment, ensure sensitive data is not improperly retained or learned from, and prove adherence to evolving legal and ethical guidelines.

How Netskope solves for this

Netskope ensures AI governance and compliance readiness through deep visibility, policy control, and real-time insights into AI usage across the enterprise.

Key capabilities include:

- **Granular Policy Enforcement:** Control how data is shared with AI tools, ensuring sensitive or regulated data is not used for unauthorized training of third-party models.
- **Real-Time Compliance Controls:** Block uploads of protected health information (PHI) to noncompliant apps or halt financial data processing in tools lacking proper certifications.
- **Regulatory Framework Support:** Ease compliance with frameworks such as the EU AI Act, NIST AI RMF.
- **Real-time Content Moderation:** Automatically filter and control harmful or discriminatory content, including hate speech, crimes, weapons, and violence.
- **Master AI Data Governance:** Secure the entire data lifecycle through automated discovery, classification, and proactive pre-deployment hardening to ensure your intellectual property remains protected and compliant.

By combining visibility, compliance intelligence, and adaptive policy enforcement, Netskope enables organizations to embrace AI innovation responsibly while meeting the ethical and regulatory demands of today and tomorrow.

The Future of AI Security

Emerging technologies and threats

As AI adoption grows and new use cases, from copilots to custom-built AI agents, become mainstream, the threat landscape is evolving just as quickly. While much of the security focus today centers around data protection and model integrity, there are two emerging areas of technology development poised to present even greater challenges in the near future.

Firstly, agentic AI systems, capable of making decisions and taking actions with minimal human oversight, are on the rise. According to Gartner, by 2028, at least 15% of daily business decisions will be made autonomously by agentic AI, up from virtually none today.[4] This shift dramatically increases the attack surface, especially if agents are granted access to enterprise systems and data through MCP or A2A (agent to agent protocol).

Secondly, physical AI, as seen in autonomous vehicles and robots, is gaining traction in industries including logistics, transportation, and manufacturing. These systems introduce real-world safety risks, where compromised or malfunctioning AI doesn't just cause data loss, but potential harm to people and infrastructure.

As AI capabilities grow more advanced and deeply embedded into everyday business operations, security leaders must establish strategic, forward-thinking governance.

Here are a few key considerations for staying ahead:

- **AI Usage Visibility:** Know which teams are building or using AI models, both open and shadow IT. Ensure central visibility and oversight without stifling innovation.
- **Data Trustworthiness:** Ensure models are trained on secure, compliant, and high-integrity datasets. Poor or tainted data leads to inaccurate, biased, or leaky outputs.
- **Autonomy and Risk Boundaries:** As agentic AI becomes more capable, define clear guardrails for autonomy. Don't wait for agents to start making high-impact decisions before governance is in place.
- **Model Life-Cycle Management:** Treat AI models like code: with version control, vulnerability scanning, access controls, and audit logs.
- **Cultural Readiness:** Security isn't just technical, it's behavioral. Educate employees and executives onAI risks, safe usage, and the evolving regulatory landscape.

The future of AI security will be defined not just by how well organizations protect against today's threats, but by how thoughtfully they prepare for what's coming next.

[4] Garner 2024 https://www.gartner.com/en/newsroom/press-releases/2024-10-21-gartner-identifies-the-top-10-strategic-technology-trends-for-2025

Prediction

Analyst firm Gartner predicts that, by 2028, at least 15% of daily business decisions will be made autonomously through agentic AI, up from 0% in 2024.

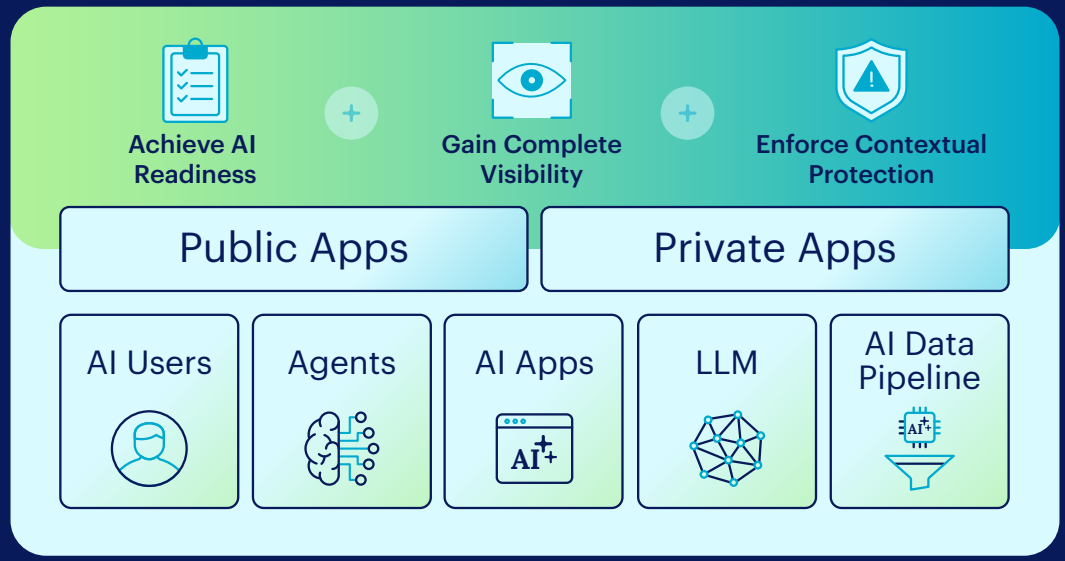
Conclusion

Secure AI end-to-end, everywhere with Netskope One

As enterprises race to adopt AI, security leaders face mounting pressure to protect sensitive data and stay ahead of new risks targeting their AI ecosystem. From a lack of visibility into AI usage to data exposure and compliance needs, we've outlined six core challenges security teams must overcome to safely enable AI across the enterprise:

- Lack of Visibility
- Understanding AI Application Risk
- AI Model Integrity
- Threats Targeting AI Systems
- Data Exposure
- Governance, Compliance, and Ethical Use

Netskope One AI Security provides a single solution to govern your AI ecosystem and protect your data. It secures users and automated agents across public SaaS, private AI tools, and agentic workflows. Combining high-performance with context-aware zero trust controls, Netskope enables organizations to unlock AI advantage and do so securely.



Research

Analyst firm Forrester found that Netskope delivers an 80% reduction in the risk of a severe breach caused by an external attack, equivalent to a $2m saving in annualized material breach costs.[5]

[5] Forrester Report: The Total Economic Impact™ of Netskope SSE
https://www.netskope.com/resources/analyst-reports/forrester-the-total-economic-impact-of-netskope-sse

About Netskope

Netskope, a leader in modern security and networking, addresses the needs of both security and networking teams by providing optimized access and real-time, context-based security for people, devices, and data anywhere they go. Thousands of customers, including more than 30 of the Fortune 100, trust the Netskope One platform, its Zero Trust Engine, and its powerful NewEdge network to reduce risk and gain full visibility and control over cloud, AI, SaaS, web, and private applications—providing security and accelerating performance without trade-offs.

Interested in learning more?

Request a demo

Resources

Securing AI with Netskope One

Mastering AI Adoption with End-to-end Security, Everywhere Blog

Netskope Threat Labs: Generative AI Cloud Threat Report

Securing Generative AI for Dummies

netskope