eBook

THALES
CYBERSECURITY

Securing AI from the Inside Out: The C-Suite's Guide to the 7 Critical LLM Risks

cpl.thalesgroup.com

Contents

Why AI Changes the Threat Model	4
The 7 Risks Every AI Team Must Understand	5
The Common Thread : Data Is Always the Target	8
A Framework for Thinking About AI Security	9
Securing the Next Generation of Data Risk	10
About Thales	11

Introduction

Every AI system runs on data, and often on the most sensitive data an organization holds. Training sets, retrieval pipelines, and agent frameworks all require deep access to internal knowledge sources.

According to the 2026 Thales Data Threat Report, 61% of organizations report that their AI applications are already being actively targeted, with sensitive data as the primary objective. The attack surface is new, but the objective is not.

AI's value depends entirely on data, and that data is now exposed on an attack surface that didn't exist 3 years ago. That attack surface is complex, comprising prompts, vector databases, agent frameworks, retrieval pipelines, and model APIs that all introduce new paths to sensitive data. Securing it was always going to be a learning curve.

This guide is for leaders who need to understand what's at stake. It outlines seven critical LLM risks, explains how they translate into real organizational exposure, and provides a framework for making confident, informed decisions about AI security.

That's what makes AI infrastructure such a compelling target: compromise the model, and you have a potential bridge to everything it can reach.

Why AI Changes the Threat Model

Although AI can come in the form of an application, AI risks are structurally different from traditional application security. Traditional security assumes clear boundaries: defined inputs, deterministic application logic, and defensive controls operating at known control points.

Large language models (LLMs), however, operate within porous and dynamic environments. Prompts, external data sources, APIs, and autonomous agents continuously interact with the model. The result is an application surface that changes based on context and data flow rather than static logic.

Security teams must also contend with how fast AI technology is changing. The AI ecosystem evolves rapidly, with new frameworks, tooling, and architectures emerging every quarter. In fact, according to the Data Threat Report, 70% of organizations cite the rate of change as their top AI security concern.

This creates several structural differences from conventional software systems:

Dynamic Inputs: Prompts and contextual data shape system behavior in real time.

Non-deterministic Outputs: Model responses vary, which complicates validation and control.

Expanding Ecosystems: Retrieval-augmented generation (RAG), plugins, and agent frameworks introduce new dependencies and integrations.

The speed of AI change makes static controls less effective and increases the risk of fragmented security decisions. This is why the OWASP Top 10 for LLMs are so important. It provides a common framework for understanding the most significant risks emerging in enterprise AI, from prompt injection to data leakage, and helps teams move from abstract concern to practical assessment.

The 7 Risks Every AI Team Must Understand

Not every AI risk is equally actionable. OWASP's Top 10 for LLMs is a comprehensive taxonomy, but several risks, including model hallucination, supply chain vulnerabilities, and agentic behavior, depend on factors largely outside an organization's direct security control. This guide focuses on the seven risks that sit squarely within the enterprise security perimeter: the data, application, and identity layers where controls can be designed, enforced, and measured.

These are the risks where security teams can act now, grouped into 3 categories: input and output integrity, data layer exposure, and infrastructure access.

Input/Output Integrity

01

Prompt injection

exploits the most fundamental characteristic of LLMs: they behave based on the instructions and context they receive. Attackers embed hidden commands in user prompts or retrieved content to override the model's intended behavior. If successful, the model may ignore safeguards, expose sensitive data, or trigger unintended actions within connected systems.

Leader's question:

If your AI system made a decision that affected a customer or partner, could you explain how it reached that conclusion and demonstrate that no one interfered with that process?

Improper output handling

occurs when AI responses are treated as trusted output without validation. If manipulated outputs feed directly into scripts, APIs, or database queries, they can trigger unintended commands or compromise downstream systems.

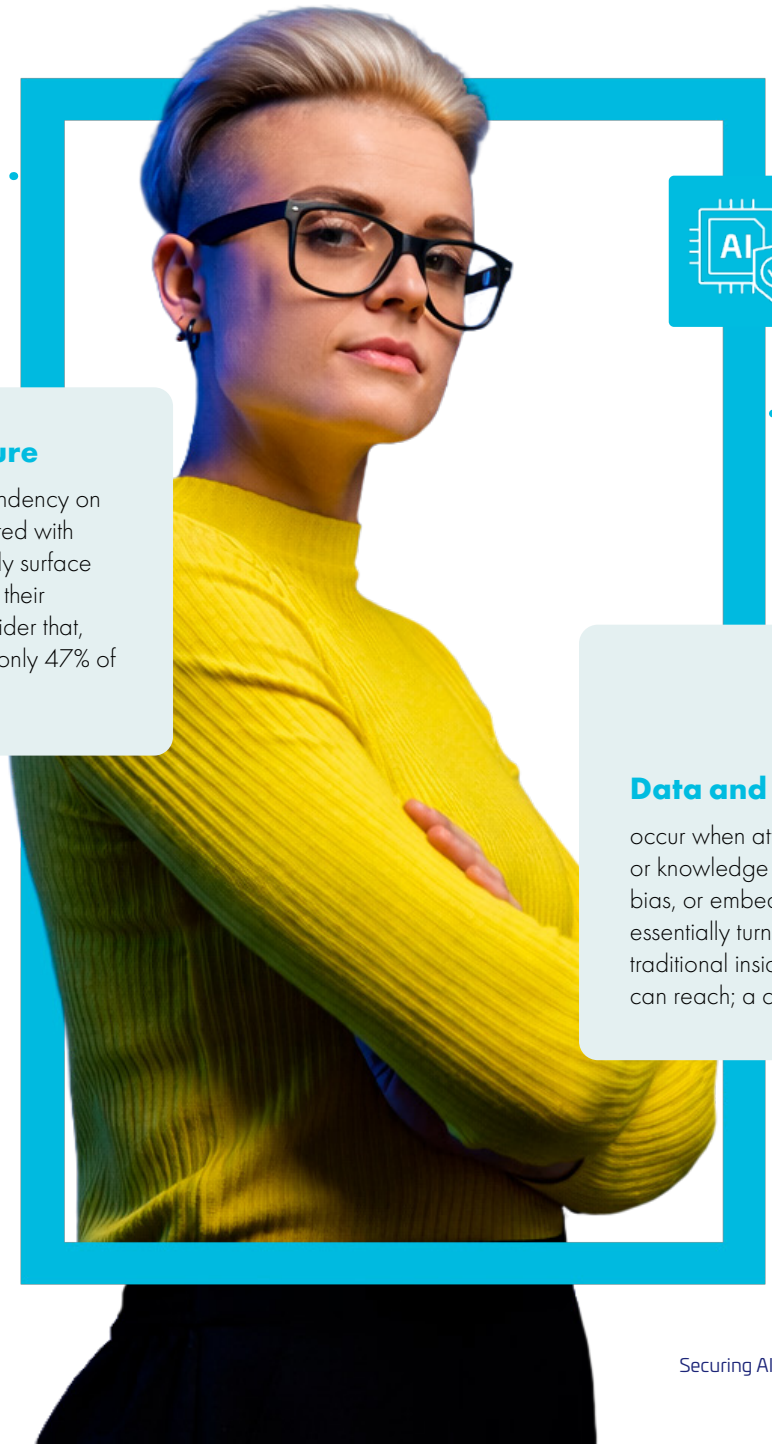
02

03

System prompt leakage

occurs when attackers extract the internal instructions that guide a model's behavior. These prompts often contain rules, guardrails, or configuration details. If exposed, they reveal how the system operates and help attackers design more effective prompt injections or bypass safeguards.

Data Layer Risks



04

Sensitive information disclosure

is the most direct consequence of AI's dependency on enterprise data. Models trained or augmented with internal knowledge sources can inadvertently surface PII, credentials, or proprietary information in their outputs. The risk compounds when you consider that, according to the 2026 Data Threat Report, only 47% of sensitive cloud data is encrypted.

05

Data and model poisoning

occur when attackers insert manipulated content into training datasets or knowledge repositories. This can alter model behavior, introduce bias, or embed hidden triggers that activate under specific conditions, essentially turning AI models into an insider threat. The difference with traditional insider threats is scale: a malicious insider affects what they can reach; a compromised model affects everything it can reach.

Leader's question:

If an auditor asked to see which data your AI systems accessed over the last 30 days, and when, could your team produce that trail with confidence?

Infrastructure and Access Risks

Leader's question:

Can you demonstrate that your AI systems operate under the same access policies and controls that govern your people and that you'd know immediately if they didn't?

06

Vector and embedding weaknesses

arise from AI's reliance on supporting infrastructure — vector databases, APIs, and compute environments — that often receive less security scrutiny than the models themselves. Weak access controls or exposed APIs can allow attackers to query vector stores directly and extract sensitive information without ever touching the model.

07

Unbounded consumption

occurs when attackers exploit the heavy computer requirements of AI models. By generating large volumes of requests or forcing expensive inference operations, attackers can degrade availability or significantly increase operational costs.

Although they are technically different, the underlying objective of each risk is to help attackers reach, manipulate, or extract sensitive data.

In this sense, AI systems behave like privileged insiders. They operate with broad, legitimate access to sensitive data sources. They act within trusted system boundaries. Their outputs appear normal until something goes wrong. This is precisely what the 2026 Thales Data Threat Report means when it frames AI as the new insider threat: not that AI is malicious by design, but that a compromised, manipulated, or poorly governed AI system has the same structural profile as a trusted employee who has been turned.

The attack surface isn't just the model. It's everything the model can touch.

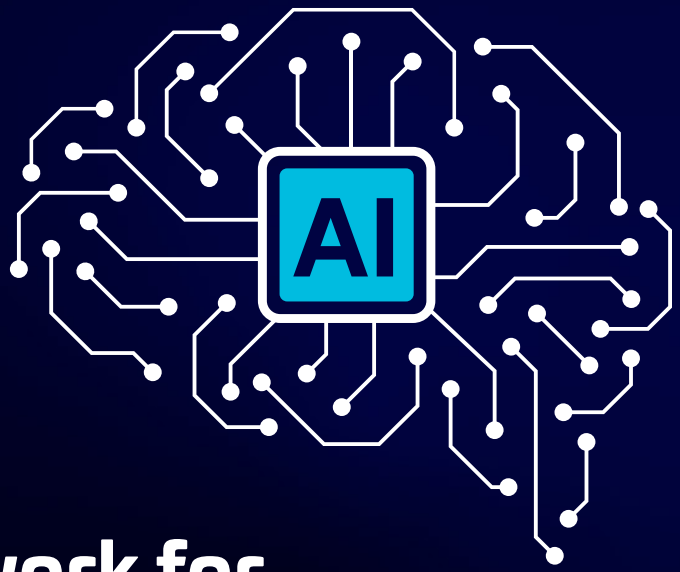
This reflects the structural reality of AI systems; models only work when they can access large volumes of enterprise data, often across multiple repositories and cloud environments. To operate effectively, they must integrate deeply with internal knowledge sources.

Integration creates opportunities for attackers. Put simply, if an attacker can compromise the AI model, they can use it as a bridge to the data that those models access.

The problem is compounded by how little visibility most organizations have into their data. According to the 2026 Data Threat Report, only 34% of organizations know where all their data resides, and only 39% can fully classify it. Without that foundation, it's impossible to reliably identify which AI interactions involve sensitive information and which require stronger protections.

Without accurate data discovery and classification, it's impossible to reliably identify which interactions involve sensitive information and, therefore, which need greater protection.

Data is everything in AI environments. It's what powers AI models, and what's most at risk. As such, security programs must start with a clear understanding of data itself. That means understanding where it resides, how it moves, and who can access it.



A Framework for Thinking About AI Security

The challenge for security leaders isn't a shortage of AI security tools; it's a shortage of coherence. According to the 2026 Data Threat Report, organizations use an average of 7 data protection tools, yet only 39% are confident they're effective. More tools without a unifying framework don't reduce risk. It redistributes it.

That's why you need a unified framework that organizes AI security across 3 core domains.

Data Security

The first domain focuses on protecting the data AI systems rely on. This includes understanding where sensitive data resides, ensuring it's encrypted, and controlling access to it.

For AI environments, these controls help limit exposure if attackers reach retrieval pipelines, training data, or vector stores. Strong data security reduces the impact of risks such as sensitive information disclosure, poisoning, and data exfiltration.

Application Security

The second domain focuses on how the AI application processes inputs and outputs. Models retrieve data and behave in response to prompts and context, making input and output validation critical.

Application-level controls help protect prompt injection, prevent system prompt exposure, and ensure model responses cannot trigger unsafe downstream actions.

Identity and Access Security

The final domain focuses on who – and what – is allowed to interact with AI systems. This includes users, services, agents, and applications that access models or the data behind them.

Strong identity and access controls ensure privileges are tightly scoped and monitored. This helps prevent unauthorized interactions, either human or machine, with AI infrastructure, retrieval systems, and sensitive data sources.

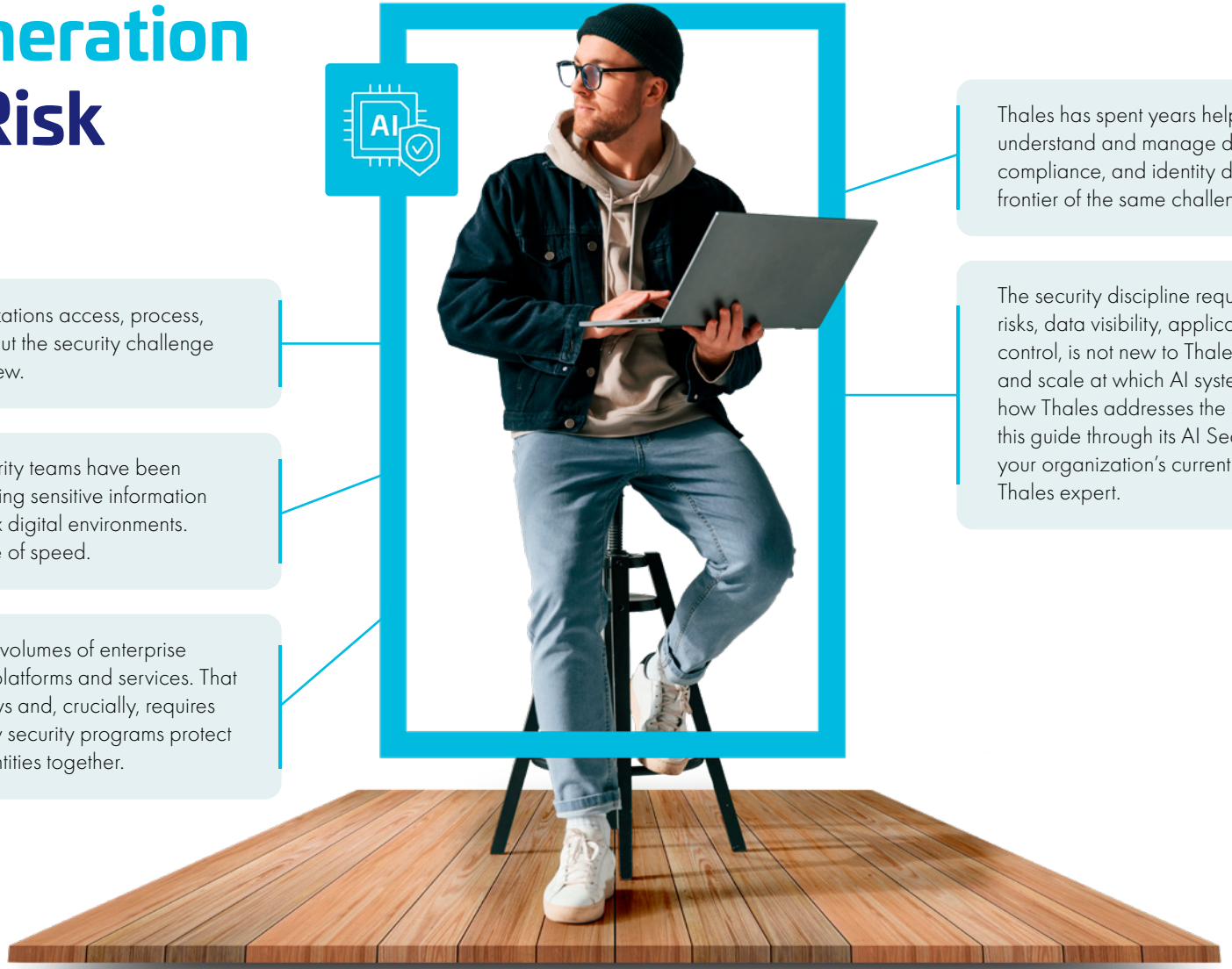
Together, these 3 domains provide a practical way to understand and manage AI risk.

Securing the Next Generation of Data Risk

AI is reshaping how organizations access, process, and generate information. But the security challenge it introduces is not entirely new.

AI expands a problem security teams have been managing for years: protecting sensitive information across increasingly complex digital environments. The difference is in the scale of speed.

AI systems interact with vast volumes of enterprise data, often across multiple platforms and services. That creates new attack pathways and, crucially, requires organizations to rethink how security programs protect data, applications, and identities together.



Thales has spent years helping organizations understand and manage data risk across cloud, compliance, and identity domains. AI is the next frontier of the same challenge.

The security discipline required to address these risks, data visibility, application integrity, and identity control, is not new to Thales. What is new is the speed and scale at which AI systems demand it. To learn how Thales addresses the seven LLM risks outlined in this guide through its AI Security Fabric , or to assess your organization's current exposure, connect with a Thales expert.

About Thales

Thales is a global leader in cybersecurity, helping businesses, governments, and the most trusted organizations in the world protect critical applications, sensitive data, identities, and software anywhere, at scale — with the highest ROI. With more than **30,000 customers**, including 58% of the Fortune Global 500, our solutions are deployed in 148 countries around the world. Through our innovative services and integrated platforms, Thales helps customers achieve better visibility of risks, defend against cyber threats, close compliance gaps, and deliver trusted digital experiences for billions of consumers every day.

THALES

CYBERSECURITY

Contact us

For contact information, please visit cpl.thalesgroup.com/contact-us

cpl.thalesgroup.com