Microsoft

Zero Trust for AI: The definitive guide to securing AI at scale

A leadership guide for the Agentic Era



AI is moving fast. Trust must keep up.

Introducing Zero Trust for AI – Microsoft's proven framework for securing agents, models, data flow and applications in your AI ecosystem.

Chapter 1: The world is changing. Security must too.

Chapter 2: The answer – Zero Trust for AI (ZT4AI)

Chapter 3: A practical guide you can use today

Chapter 4: The ZT4AI Architecture – Extending Zero Trust across the AI Lifecycle

Chapter 5: Securing AI using Microsoft Security platform

Chapter 6: How Microsoft does it – SFI as the proof point

Chapter 7: Security as a business enabler

Chapter 8: A four-stage operating model for ZT4AI

Scale AI with confidence, not compromise.

The world is changing. Security must too.

The Frontier Firm is here

AI is no longer a future investment, it is today's operational reality. Every industry is feeling the shift. Organisations are deploying copilots, building autonomous agents and embedding AI into core workflows at a pace we've never seen before.



More than **80% of leaders** say they're already using agents or will be using agents within the next year. IDC projects **1.3 billion agents by 2028**, roughly the population of India. That might be an underestimate because every human worker will have multiple agents acting, deciding and accessing data on their behalf.

This moment is giving rise to what we call the Frontier Firm: where human ambition and intelligent agents come together to help people do things they simply couldn't do before. That's an extraordinary opportunity. But it comes with a profound security responsibility.

The scale changes everything: how **we work**, how **we secure** and how **we build** for the future.

AI Introduces new security challenges that keep security decision makers up at night.

The rise of AI doesn't just accelerate existing threats. It creates entirely new ones. Security leaders consistently identify four challenges they cannot afford to ignore:

1. Agent Sprawl & resource access

82%

of leaders expect to use agents in the next 12–18 months[1]

AI agents act autonomously, accessing systems and data at machine speed – often with excessive permissions.

2. Data leaks & oversharing

80%

of leaders cite sensitive data leakage as their top concern

AI systems reason over sensitive data, creating new vectors for leakage and policy violations.

3. New AI threats & vulnerabilities

88%

of organisations are concerned about indirect prompt injection attacks

Prompt injection, data poisoning, model manipulation and adversarial inputs are now active threat categories.

4. Regulatory compliance

55%

of leaders lack understanding of how AI will be regulated

AI governance requirements are multiplying – EU AI Act, NIST AI RMF, ISO 42001 – and most organisations aren't ready.

1. 82% – *Source: Microsoft Work Trend Index Survey 2025* (2025: The year the Frontier Firm is born) 80% - Source: Securing AI: Navigating risks and compliance for the future | The Microsoft Cloud Blog

And attackers aren't standing still. AI-automated phishing is now 4.5× more effective. Malware, deepfakes and adversarial campaigns can be generated and deployed at unprecedented speed. Organisations defending against these threats are already stretched – many running 40, 60, even 80 disconnected security tools that create blind spots and slow response.

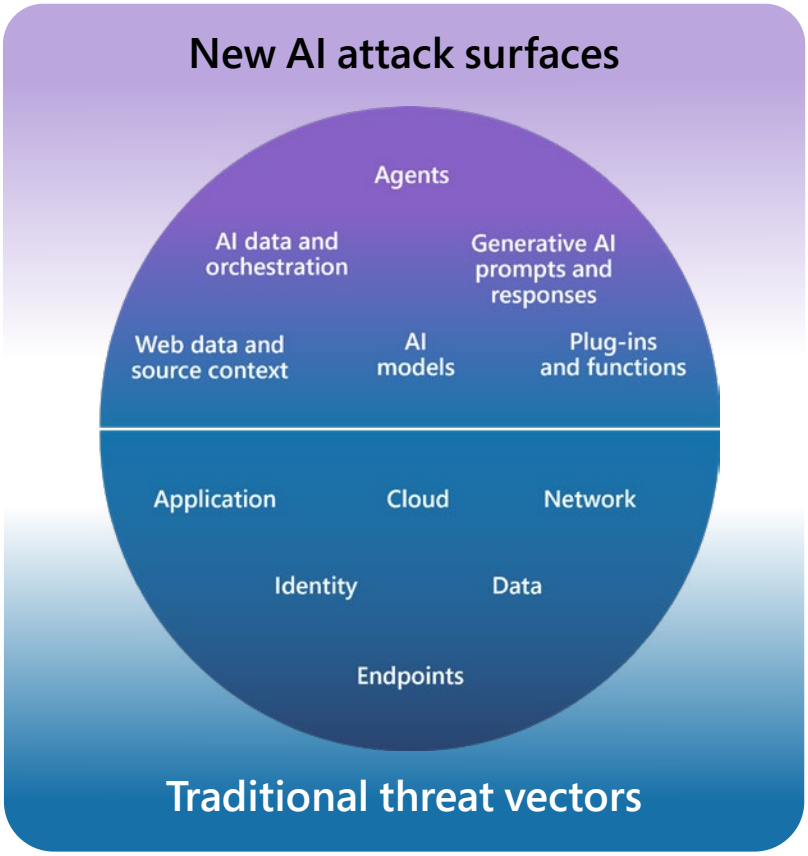
The conclusion is clear: traditional security architectures are not designed for systems that act autonomously, reason over sensitive data and continuously evolve. The strategies that protected application, cloud, network, identity and data environments remain important, but they are no longer sufficient on their own.

Thriving in a landscape of threats

Cybercrime is growing rapidly, with attacks increasing in speed, scale and sophistication. Nation-state actors are increasingly working alongside financially motivated criminals, creating threats that are more complex and difficult for defenders to counter.

Artificial intelligence is accelerating this shift. Attackers are using AI to reduce the cost, time and expertise required to launch attacks, allowing malicious campaigns to scale faster than ever before. Malware, deepfakes and other attack techniques can now be generated and deployed with unprecedented speed.

At the same time, organisations must secure technologies that did not exist only a few years ago – AI prompts and responses, the data pipelines that power models and agents, plugins and integrations that expand systems and autonomous agents capable of acting at machine speed.



Each security challenge introduces a new attack surface – one that adversaries can probe, exploit or attempt to poison.

This is why the current moment represents a fundamental shift in cybersecurity.

In a world increasingly shaped by artificial intelligence and autonomous agents, security must evolve just as rapidly as the technologies it seeks to protect.

Zero Trust for AI

What is Zero Trust for AI?

Zero Trust for AI (ZT4AI) is the application of Zero Trust security principles – verify explicitly, use least privilege access and assume breach – to every aspect of AI systems, agents and workflows.

This means:
- Every identity accessing AI – human or non-human – is continuously validated
- Every AI agent and application operates with only the minimum permissions required
- Every interaction, prompt and response is treated as a potential risk
- Every data flow through AI pipelines is governed, classified and protected

ZT4AI is the architecture and foundation for practical, scenario-driven implementation through Microsoft's Security for AI (S4AI) – operationalising it with products, experiences, controls and implementation guidance customers can apply immediately. Together, they deliver a single, unified approach to securing AI innovation.

Why Zero Trust is the right framework for AI

Zero Trust principles were designed for exactly the kind of environment AI creates – one where the perimeter is gone, identities are non-human, data moves at machine speed and breach is assumed from the start.

Here is how you can use the Zero Trust principles to address the core AI security challenges:

1. Business Enablement
ZT4AI embeds guardrails that allow AI initiatives to move faster, scale responsibly and deliver measurable business impact.

3. Use Least-Privilege
Govern agent permissions with just-in-time access, scoped roles and bounded tool usage.



2. Verify Explicitly
Continuously validate every agent identity, model request and API call, not just once at login.

4. Assume Breach
Detect and contain prompt injection, data poisoning and adversarial inputs before they spread.

Practical guidance you can use today

From strategy to execution: let's start your ZT4AI journey

Zero Trust for AI is not theoretical. Microsoft is delivering practical, scenario-based tools that help organisations move from 'Where do we start?' to 'What do we do next?' in a secure, programmatic and scalable manner.

"It's not just about detecting threats. Microsoft Defender XDR unifies our security environment, so everything works together seamlessly."

Ulrich Koinig
Cyber Defence Manager,
OMV Aktiengesellschaft

Read the full story

"We turned to Microsoft Security Copilot because it integrates with our Microsoft stack. Analysts can ask questions in human language and get context they can act on, which speeds up the handoff between the SOC and our IT team."

Kalunyu Sopha
Security Manager, Toyota Leasing Thailand

Read the full story

"Our adoption of the integrated security platform within Microsoft 365 E5 led to a significant reduction in mean time to detect and security incidents response, along with cost saving from tool consolidation."

Timothy Torres
Chief Security Officer, TriNet

Read the full story

"The Microsoft security stack is more than technology. It contributes to Ford's business in moving faster against cyberthreats and building a more secure future."

Weston Maggetti
Platform Manager, Ford Motor Company

Read the full story

The Zero Trust Workshop

Customers have consistently told us that they see Zero Trust as a strategic foundation for how they approach and run a modern security practice. However, customers have also shared that they need help and guidance understanding how their security products could help them achieve a Zero Trust vision, and how they can measure how far along they are on their journey to implement Zero Trust.

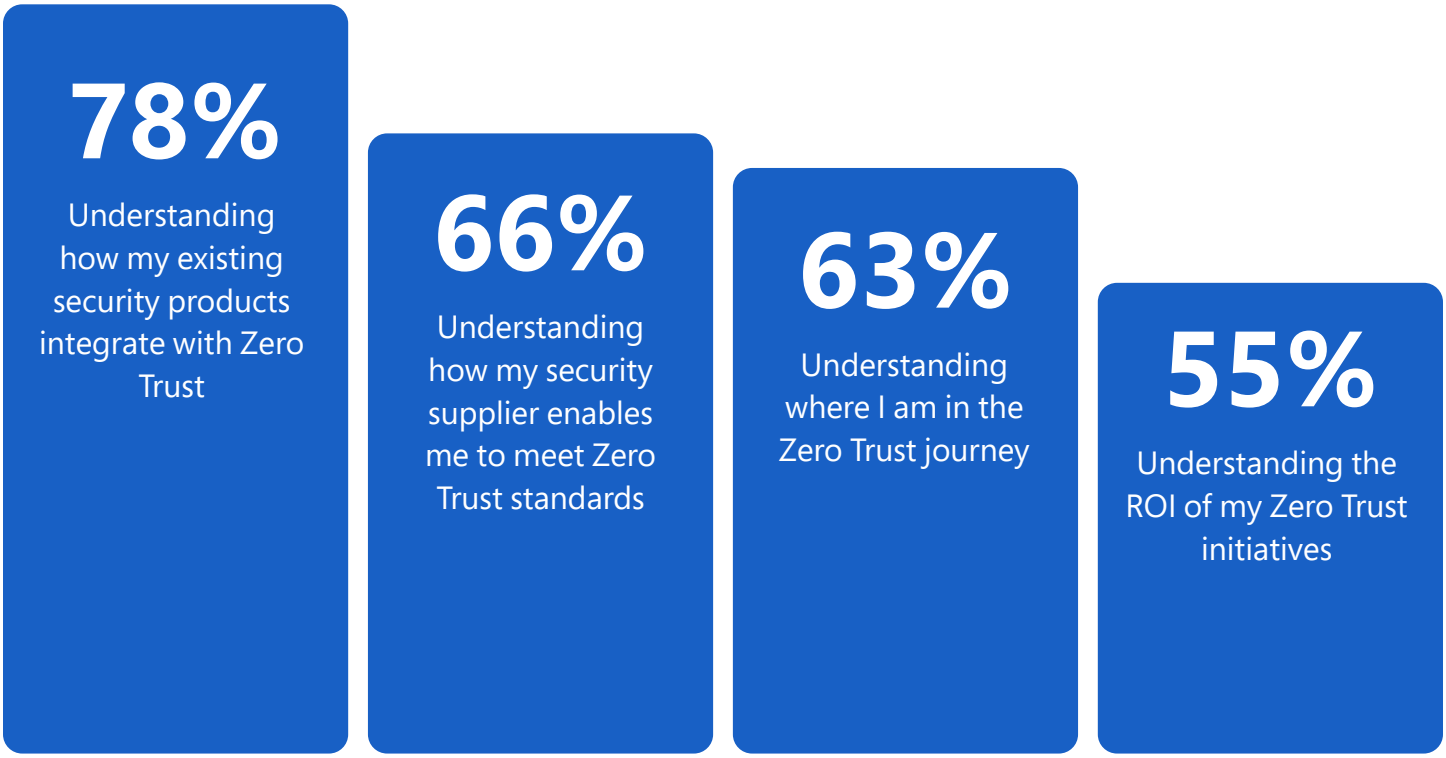


Figure 1. Top areas customers ask for help with in advancing their Zero Trust journey. (source: Microsoft survey)

The **Zero Trust Workshop** is a publicly available resource designed to help organisations move from security intent to measurable action. Instead of abstract principles, it provides a prescriptive, scenario based path to strengthen security while enabling the business outcomes. With the addition of the new AI pillar, the workshop now covers 700 security controls across 116 logical groups and 33 functional swim lanes. Customers can now apply Zero Trust to real-world AI scenarios, helping them deploy and scale AI with confidence while managing new risks such as over-privileged agents, data exposure and governance gaps.



The ZT Workshop includes these AI security scenarios:

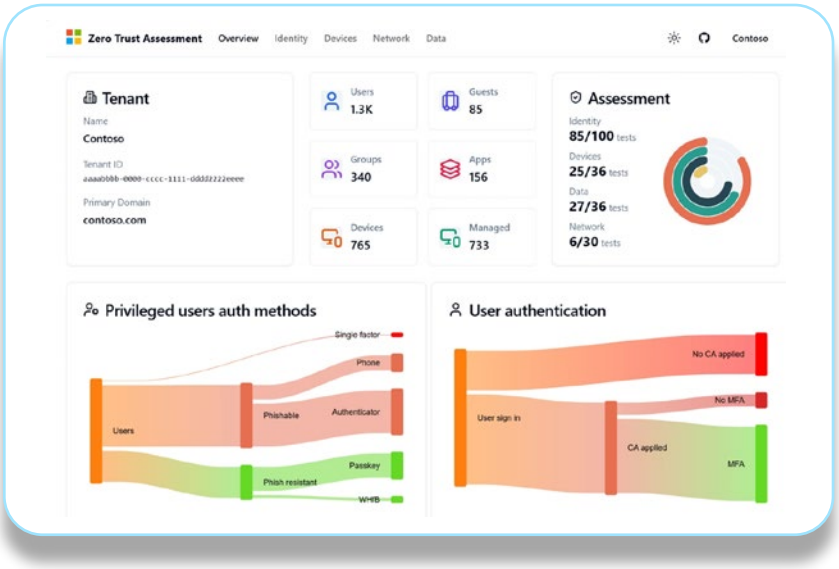
- **Register all Agents** – Inventory and govern every agent in your environment
- **Secure AI Auth & Access** – Apply strong authentication and least-privilege to AI identities
- **Secure AI Network Access** – Control and segment AI system network paths
- **Secure AI Data Access** – Classify and protect data accessed by AI models and agents
- **Build Security Agents** – Embed security into agentic development workflows
- **Security Detection & Response for AI** – Enable monitoring, alerting and response for AI specific threats

Run the workshop yourself: https://aka.ms/ZTWorkshop
Request delivery: https://aka.ms/ZTworkshop/request

Microsoft Zero Trust Assessment

As AI agents become more capable, the stakes around data and network security have never been higher. Agents that are insufficiently governed can expose sensitive data, act on malicious prompts or leak information in ways that are difficult to detect and costly to remediate. Data classification, labelling, governance and loss prevention are essential controls. So are network-layer defences that inspect agent behaviour, block prompt injections and prevent unauthorised data exposure.

Yet, manually evaluating security configurations across identity, endpoints, data and network controls is time consuming and error prone. That is why we built the **Zero Trust Assessment** to automate it. The Zero Trust Assessment evaluates hundreds of controls aligned to Zero Trust principles, informed by learnings from Microsoft's Secure Future Initiative (SFI). The Zero Trust Assessment transforms the time consuming, error prone process of manually checking security configurations into an automated, intelligence-driven experience.



Tests are drawn from:

- **Industry standards** – NIST, CISA and CIS
- **Microsoft's own SFI baselines** – the same controls protecting Microsoft's internal infrastructure
- **Real-world insights** from thousands of security implementations

See a demo: https://aka.ms/zerotrust/demo
Run your assessment: https://aka.ms/zerotrust/assessment

Patterns & Practices **Actionable engineering guidance**

To operationalise ZT4AI at the engineering and implementation level, Microsoft has introduced a new library of Patterns & Practices, practitioner-focused guidance that extends Zero Trust from architecture into day-to-day engineering and operations.

AI observability: Making AI systems observable and accountable through end-to-end logging, traceability and monitoring that enable oversight, incident response and trust at scale.
https://aka.ms/SFI_AIObservability

Identify risk for autonomous agentic AI systems: Case study outlining top risks of agents and why we need a layered approach
https://aka.ms/SFI_AgentsDefenseinDepth

Secure autonomous agentic AI systems: Actionable guidance teams can apply today – covering agent lifecycle management, identity and access controls, policy enforcement, safe tool usage and operational guardrails – to reduce risk without stalling innovation.
https://aka.ms/SFI_SecuringAgenticSystems

Defend against indirect prompt injection attacks: How it works, why traditional mitigations fail and how a defence-in-depth approach – spanning input handling, tool isolation, identity, memory controls and runtime monitoring – can meaningfully reduce risk.
https://aka.ms/SFI_IndirectPromptInjection

Remove legacy systems that risk security: Microsoft implemented a comprehensive cleanup initiative under the Secure Future Initiative (SFI) to eliminate legacy and unmanaged tenants.
https://aka.ms/SFI_Removelegacysystemsthatrisksecurity

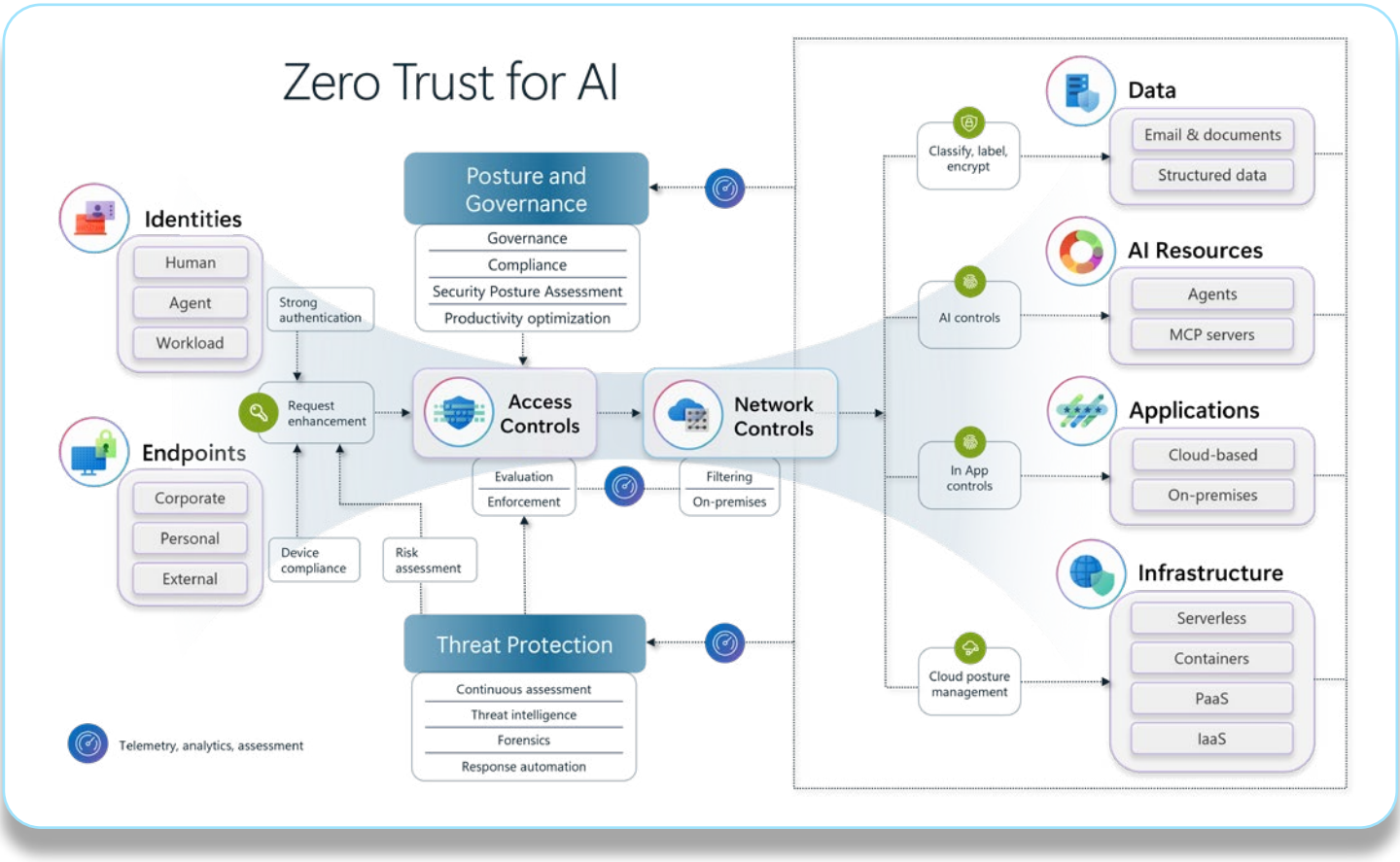
Threat Modelling: Threat modelling for AI systems is an ongoing engineering mindset rather than a one-time checklist. It begins with identifying what must be protected, understanding how the real system behaves end to end, modelling intentional misuse and designing architectural mitigations that constrain failure.
https://aka.ms/SFI_ThreatModeling

These patterns help organisations build AI systems that are secure, resilient and worthy of trust – bridging the gap between architectural principles and operational reality.

The ZT4AI architecture – Extending Zero Trust across the AI lifecycle

A new reference architecture for the Agentic Era

Microsoft's new **Zero Trust for AI reference architecture** extends an existing Zero Trust architecture across the full AI lifecycle, covering identities, agents, data, models, infrastructure, networks, endpoints and threat protection.



This architecture provides a common framework for security, IT and engineering teams, helping organisations understand:

- Where controls apply across the AI stack
- How trust boundaries shift with agentic AI
- Why defence-in-depth is essential for AI-powered workloads
- How to limit blast radius when failures occur

Identities – Human and Agent

Every access request – from a human user, an AI agent, a workload identity or an MCP server – must be authenticated with strong authentication and enforced through least-privilege policies based on compliance status and behaviour.

Endpoints – Devices and AI Execution Environments

Managed devices, Bring Your Own Device (BYOD), partner systems, IoT and hybrid cloud environments all expand the attack surface. Access is granted only after verifying device health, compliance and trustworthiness – including the execution environments where agents run.

Applications – Including Copilots, Plugins and Agents

Secure cloud-based and on-premises applications. Apply modern authentication and API protection. Control plugins, functions and integrations that extend AI system capabilities and introduce new attack surfaces.

Data – Classify, Label, Protect

Data is the primary asset AI touches. Apply data classification, context-aware access and data protection – including DLP and data lineage – across AI pipelines to prevent oversharing and leakage.

Networks – Segment and Encrypt

Micro-segmentation, end-to-end encryption, continuous monitoring and threat protection prevent unauthorised access and lateral movement – including from compromised agents or model endpoints.

Infrastructure – Serverless, Containers, PaaS, IaaS

Protect workloads and configuration. Apply cloud posture management and continuous assessment to AI infrastructure, including Azure AI Foundry deployments and third-party AI services.

Security Operations – Detection and Response for AI

Enable comprehensive visibility and automated threat response across the environment. Translate proven incident response principles to AI drive systems that address new harm types, telemetry gaps and emerging AI-specific attack patterns.

AI Resources

Use models, agents, copilots, tools and automation to operate across applications, networks and data. Properly defining and governing these systems ensures appropriate controls are applied and emerging risks are reduced.

Securing AI using Microsoft's Security Platform

One Unified Security Platform for the Agentic Era

ZT4AI is operationalised through Microsoft's AI-first, **end-to-end security platform** – integrated product families that share signal, context and intelligence across the entire environment:

Microsoft Entra – Secure access for every user and every agent. Apply least-privilege and continuous evaluation to identities at human and machine scale. Govern agent service principals and workload identities.

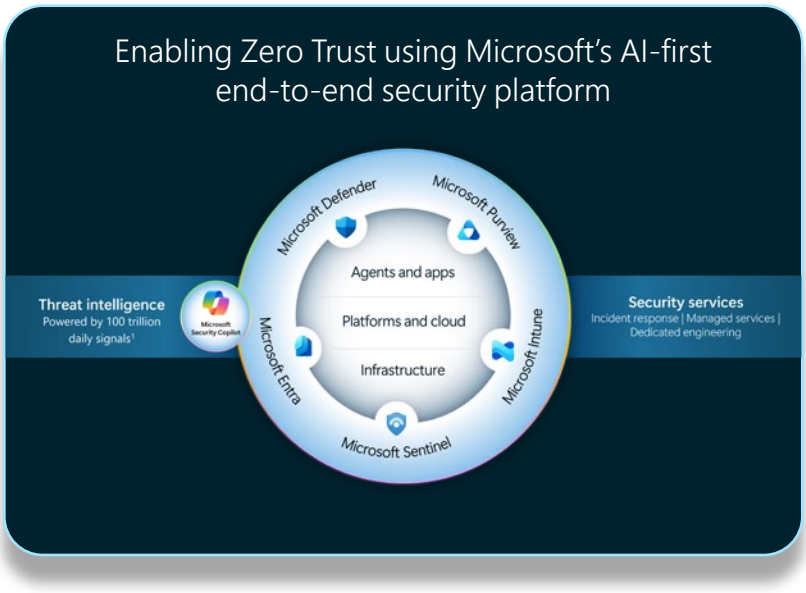
Microsoft Purview – Classify, label and protect data wherever it lives – across cloud, apps, devices and AI systems. Prevent oversharing and data leakage in AI pipelines. Enable compliance and data lineage at scale.

Microsoft Defender – AI-powered threat protection unifying prevention, detection and response across endpoints, identities, cloud workloads, email and AI systems. Detect and respond to prompt injection, model abuse and adversarial inputs.

Microsoft Intune – Ensure every device that connects to your organisation and your AI systems is configured safely and managed consistently.

Microsoft Sentinel – The platform layer that connects it all. Unify signals and context so security actions are informed, coordinated and increasingly autonomous – with AI-specific detection and response capabilities.

Microsoft Security Copilot – Runs across the entire platform, giving defenders intelligence, speed and autonomy to operate at the pace of today's threats. Transforms complex signals into clear insights, enabling faster knowledge discovery, investigation and decision-making.



Microsoft 365 Copilot
Secure and govern Copilot usage – control data access, classify sensitive content, monitor for oversharing.

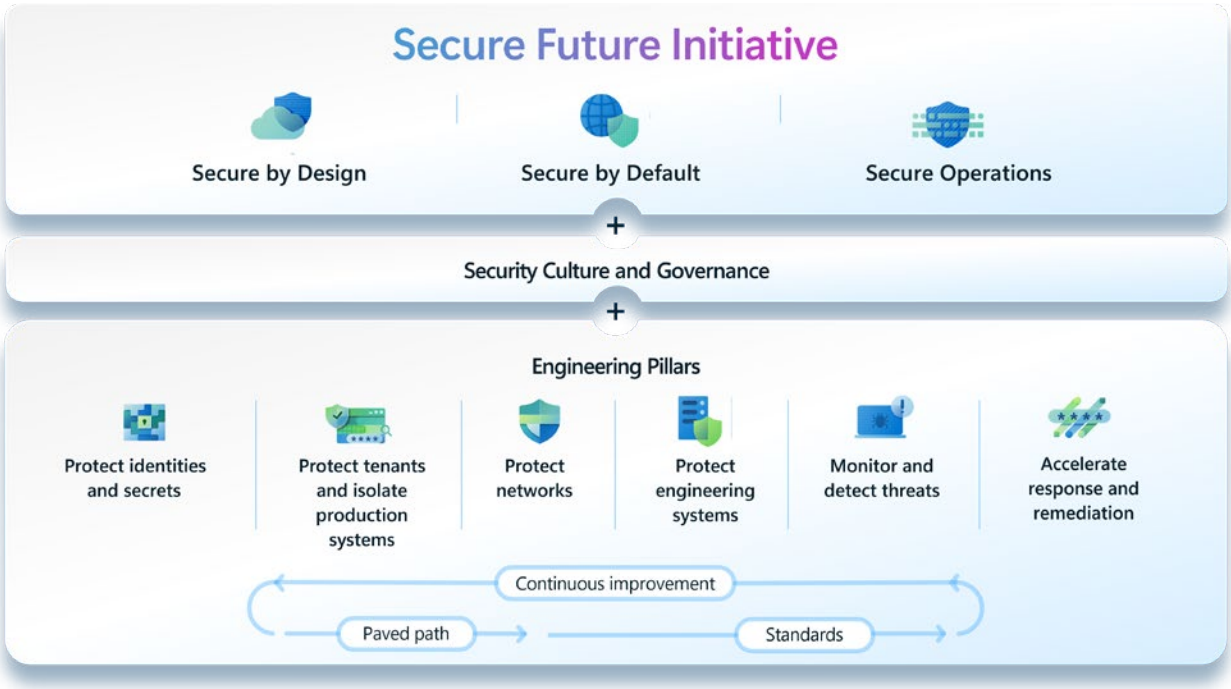
Microsoft Foundry
Build secure, trustworthy AI agents with embedded Zero Trust controls across identity, data and infrastructure.

Microsoft Copilot Studio
Create secure agents using little to no code – with least-privilege access and policy enforcement built in.

Third-party AI apps & agents
Protect all AI apps regardless of origin – apply consistent governance, monitoring and access controls.

How Microsoft does it – SFI as the proof point

Security at scale, proven internally



Microsoft's commitment to Zero Trust for AI isn't theoretical – it's been operationalised at every layer of how we build. Through the **Secure Future Initiative (SFI)**, security is now baked into how we design, how we ship and how we operate – representing the **largest cybersecurity engineering investment in Microsoft's history**, with more than **35,000 engineers** dedicated to this work.

SFI is guided by three core principles that directly inform ZT4AI:

- **Secure by Design** – Embed security into the earliest stages of AI development
- **Secure by Default** – Ensure the safest configuration is the standard, not the exception
- **Secure Operations** – Treat security as continuous, with ongoing monitoring and adaptation

The SFI engineering pillars – protecting identities, isolating production systems, securing networks, protecting engineering systems and accelerating response – are the same controls Microsoft is sharing with customers through ZT4AI guidance. When we say "how Microsoft does it," we mean it: **our internal best practices become your implementation playbook.**

Recognised Leadership

Microsoft has been named a Leader in The Forrester Wave™ for Zero Trust Platforms, Q3 2025, ranking highest in the strategy category – recognised for its proactive security architecture, AI-driven capabilities and mature partner ecosystem that enables Zero Trust at global scale.

The Forrester Wave™ Zero Trust Platforms, Q3 2025

Forrester names Microsoft a Leader, ranking highest in the strategy category



The Forrester Wave™ is copyrighted by Forrester Research, Inc. Forrester and Forrester Wave™ are trademarks of Forrester Research, Inc. The Forrester Wave™ is a graphical representation of Forrester's call on a market and is plotted using a detailed spreadsheet with exposed scores, weightings and comments. Forrester does not endorse any supplier, product or service depicted in the Forrester Wave™. Information is based on best available resources. Opinions reflect judgment at the time and are subject to change.

Security as a business enabler

The shift that changes everything

At its core, Zero Trust for AI reflects a fundamental shift in how security supports innovation. Rather than slowing AI experimentation, ZT4AI embeds guardrails that allow AI initiatives to **move faster, scale responsibly and deliver measurable business impact.**

ZT4AI extends this momentum into the AI era – providing a **unifying security frame** for all AI-related customer conversations, simplifying sales and partner motions and embedding Microsoft's architecture directly into customer AI transformation plans.

What customers gain

With Zero Trust for AI, your organisation gets:

- **A prescriptive workshop** to move from insight to execution with your team
- **A structured assessment** to understand AI security readiness and prioritise gaps
- **A clear architectural model** for securing AI systems end-to-end, from agents to models to data pipelines
- **Practical patterns and practices** for engineering teams to build securely today
- **Reduced risk from over-privileged agents, data leakage, model manipulation and prompt injection**
- **Confidence to scale AI** without compromising security, compliance or trust
- **Board-ready reporting** with executive summaries that align CISOs, CIOs and business leaders

A Four-stage operating model for ZT4AI

Securing AI with Zero Trust can be organised into four operational stages – giving your team a clear, sequenced path from foundation to full lifecycle governance:

1. Prepare
Establish a Zero Trust foundation across identity, access and data. Define classification, lineage and governance policies. Ensure sensitive information is handled consistently and access is enforced through least-privilege controls – before AI systems go into production.

2. Discover
Maintain continuous visibility into AI activity – including prompts, responses and supporting data. Continuously assess security posture across AI applications, models, orchestration layers and development pipelines to identify exposure and operational risk.

3. Protect
Apply data protection, endpoint controls and workload security to prevent misuse, data leakage and adversarial threats. Implement safety controls to manage content risk and maintain operational integrity across your AI estate.

4. Govern
Align AI operations with regulatory, legal and internal policy requirements. Enable lifecycle management, audit readiness, eDiscovery and communications governance. Support compliance with **EU AI Act, NIST AI Risk Management Framework, ISO 42001, ISO 23894** and internal codes of conduct.

Your ZT4AI journey starts here

Three steps to get started today

1. Run the Zero Trust Workshop with the AI Pillar
2. Assess your AI security readiness
3. Apply Patterns & Practices to your engineering teams

© Copyright Microsoft Corporation 2026. All rights reserved