NeuralTrust

# The State of AI Agent Security 2026

# Foreword

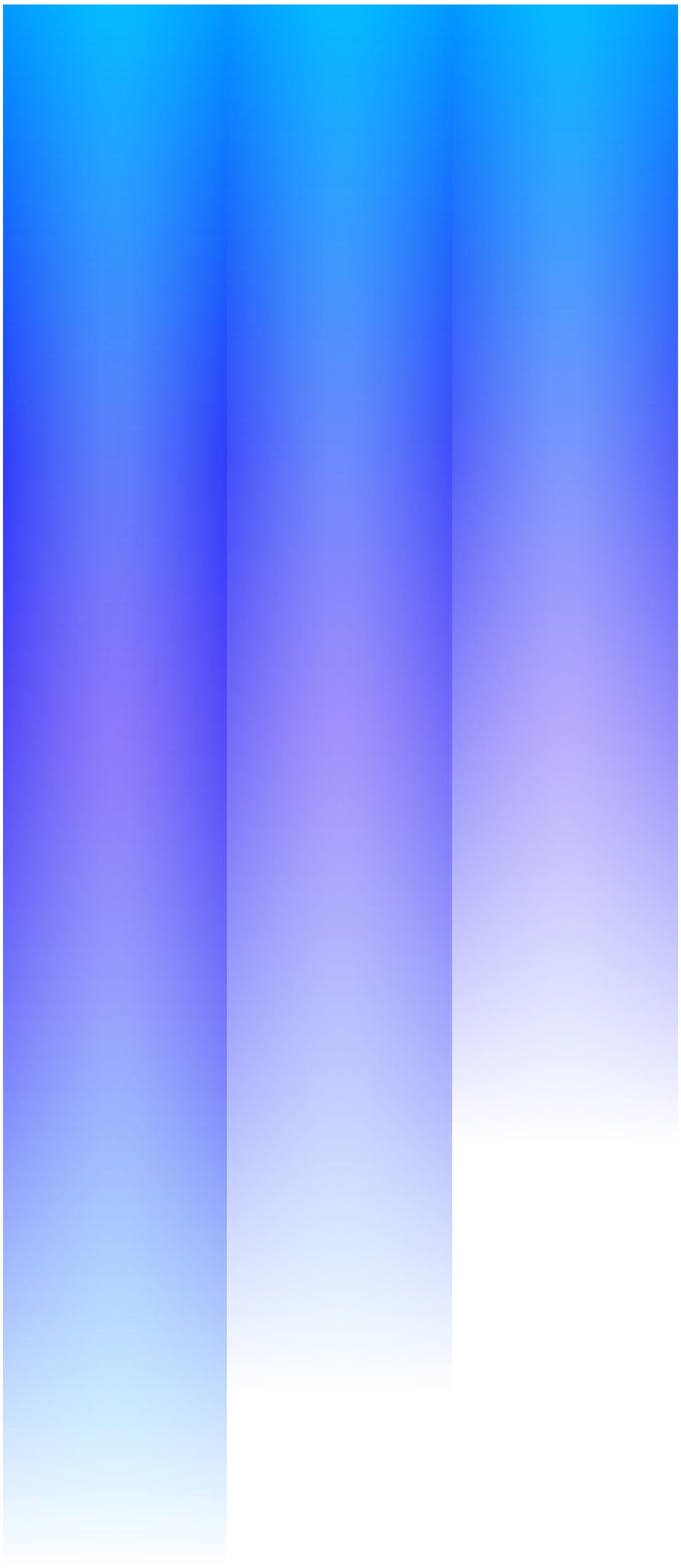
We are witnessing a turning point in enterprise technology. In just a few years, AI has evolved from experimental chatbots to autonomous agents capable of making decisions, retrieving information, and taking actions across live systems. These agents promise efficiency, creativity, and speed, but they also introduce a new and complex security challenge that few organizations are fully prepared for.

At NeuralTrust, we spend every day helping enterprises navigate this transformation responsibly. What we've learned is clear: AI adoption has moved faster than AI security.
Traditional controls that are built for static applications and human-triggered events just can't keep pace with systems that reason, adapt, and act autonomously.

This report, The State of AI Agent Security 2026, reflects the collective voice of over 160 CISOs and security leaders globally. It captures a moment where innovation is racing ahead of regulation, and where the need for trust, visibility, and accountability has never been more urgent.
The next phase of AI will not be defined by who deploys the most agents, but by who secures them best.
My hope is that this study sparks an honest conversation about what readiness really looks like and how, together, we can build a future where intelligence and integrity advance hand in hand.

— Joan Vendrell, CEO NeuralTrust

# Executive Summary



**The State of AI Agent Security 2026**
Enterprises are deploying AI agents faster than they can secure them. 72% have implemented or are scaling AI agents, yet only 29% report comprehensive security controls.

This readiness gap defines the current AI landscape: rapid adoption, limited governance, and minimal visibility into agent behavior.
While 65% monitor activity, few perform red-teaming or behavioral analysis.

73% of CISOs are very or critically concerned about AI risks, but only 30% have mature safeguards.
Early incidents confirm the threat: 1 in 5 organizations report at least one AI agent–related breach, mostly from prompt injection or data exposure.
40% estimate financial losses of $1–10 million, and 13% expect more than $10 million, comparable to ransomware-level impact.
Control maturity remains fragmented.
Most rely on monitoring (42%), access control (38%), and DLP (31%), while few employ adversarial testing (19%) or supply chain validation (16%).A quarter of enterprises have no AI-specific controls at all.
NeuralTrust's maturity model places 46% of organizations in the Reactive tier and 29% in Managed, with fewer than 10% reaching Proactive governance.
Europe leads in readiness, driven by the EU AI Act, DORA, and NIS2, while North America leads in experimentation.
By 2028, one in three enterprises will operate over 500 AI agents, and by 2030, more than half will.
At the same time, 80% will fall under AI-specific regulation, and three-quarters will employ dedicated AI security specialists.
The message is clear: adoption has outpaced defense.
Enterprises must map AI assets, monitor continuously, test proactively, embed governance, and prepare for global compliance.
AI assurance will soon define leadership, those who act now will lead in both innovation and trust.

# Table of Contents

# Methodology

This report is based on a global survey of 160 CISOs, Chief Data Officers, and cybersecurity leaders, conducted between August and September 2025.

Survey design & instrument

- Structured online questionnaire with 12 multiple-choice questions focused on AI agent adoption, risk perception, controls, incidents, and business impact.
- No personally identifiable information collected; responses are anonymous.

160+ responses

>500$ million ARR

August - September 2025

43% North America
36% Europe
12% Asia-Pacific
6% Latin America
3% Middle East & Africa

# Section II
# The New Frontier of AI Agents

Enterprises are moving rapidly from AI experimentation to AI execution. What began as LLM pilots is now scaling into autonomous AI agents embedded across business operations.

This section explores how adoption has evolved, where organizations stand today, and what's next:

- The Rise of AI Agents
- Where Enterprises Stand Today
- The Road Ahead

# The Rise of AI Agents

AI adoption has entered a new phase. What began as LLM pilots for internal productivity is now expanding into autonomous agents operating at enterprise scale. According to The State of AI Agent Security 2026, nearly every organization has already experimented with large language models, but the shift toward AI agents capable of taking action is in motion.

**70%** of organizations now use LLMs for internal productivity, showing that AI has become embedded in everyday workflows,  for example accelerating research, reporting, and code generation across teams.

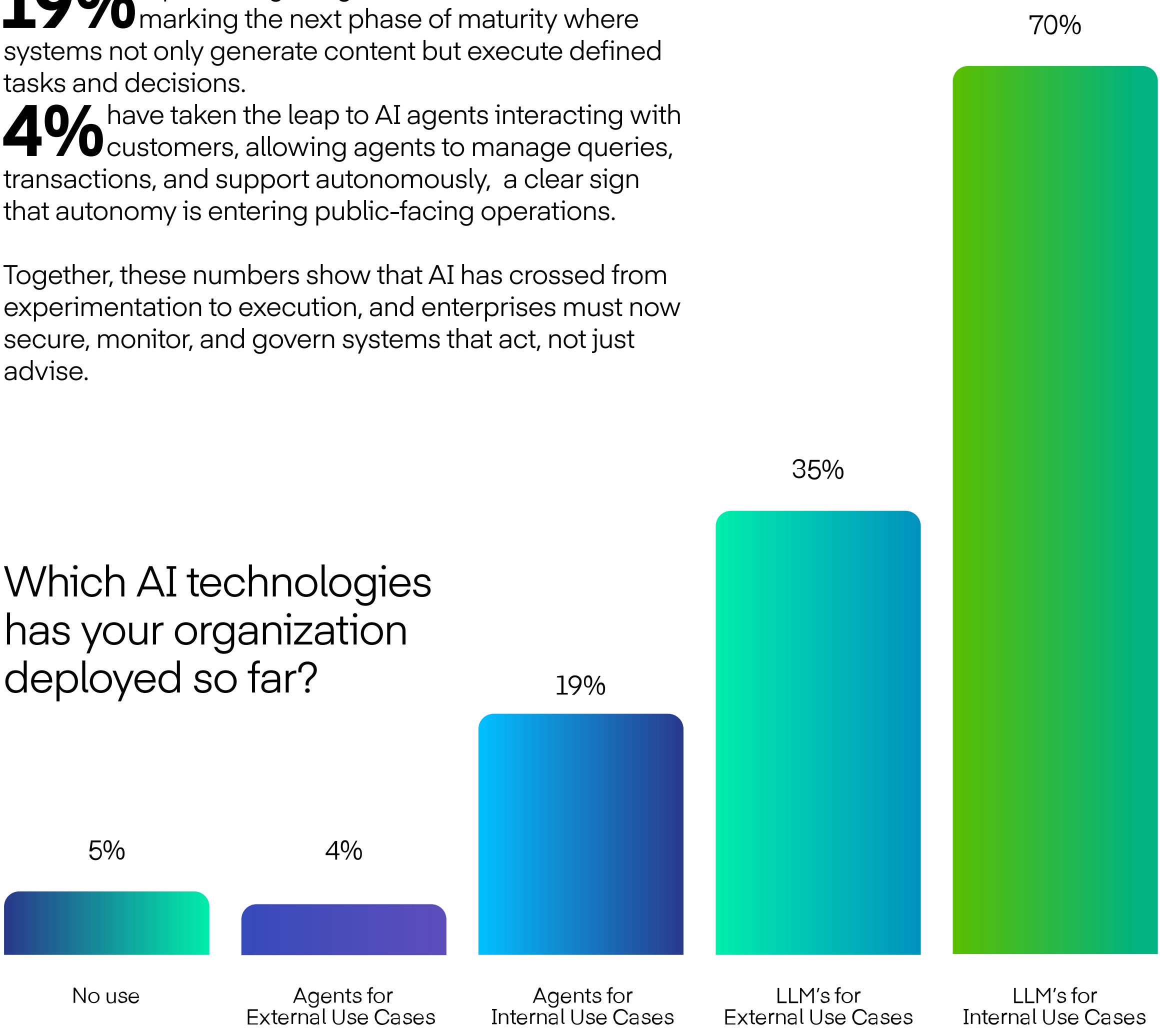
**35%** are already deploying LLMs in customer-facing roles, signaling growing trust in AI to interact directly with clients and shape brand experience.

**19%** report using AI agents for internal workflows, marking the next phase of maturity where systems not only generate content but execute defined tasks and decisions.

**4%** have taken the leap to AI agents interacting with customers, allowing agents to manage queries, transactions, and support autonomously,  a clear sign that autonomy is entering public-facing operations.

Together, these numbers show that AI has crossed from experimentation to execution, and enterprises must now secure, monitor, and govern systems that act, not just advise.

## Which AI technologies has your organization deployed so far?



Source: The State of AI Agent Security 2026, Q2 "Which AI technologies has your organization deployed so far?" (n = 164 respondents)

# Where Enterprises Stand Today

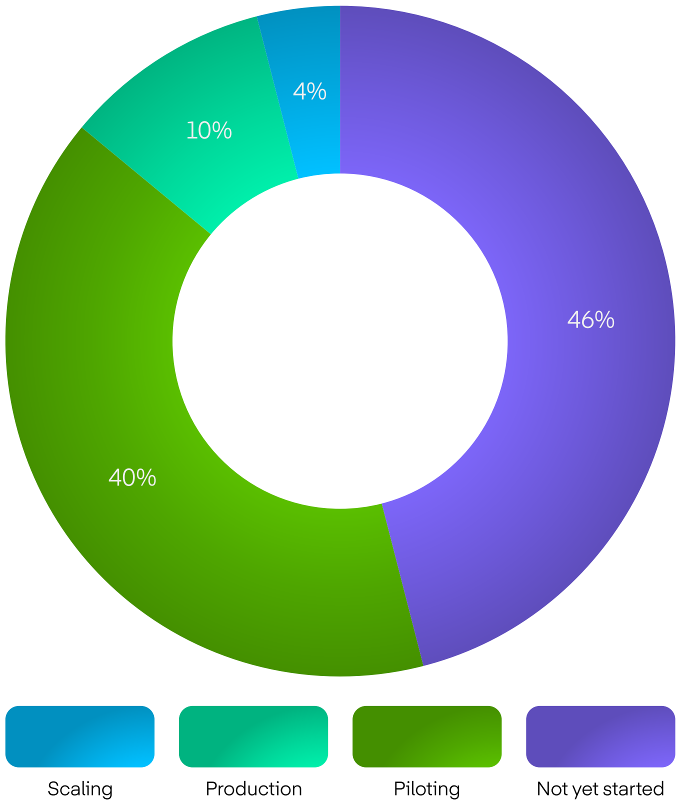
AI has moved from isolated pilots to enterprise-wide deployment. Across industries, organizations are transitioning from testing to scaling, integrating AI agents directly into operations, compliance, and customer functions.

Only **4%** of respondents report **scaling AI agents** across multiple departments, while **10% have fully deployed them in production.**
A much larger share **(40%)** remain in pilot or proof-of-concept stages, and nearly **half (46%) have not yet started.**
This distribution shows that while interest in AI agents is widespread, true operational maturity is still incipient.

Enterprises are experimenting and validating business use cases but have yet to embed AI agents into critical systems at scale.

The State of AI Agent Security 2026 asked respondents to select which stage of AI Agent adoption was already in use at their organization and the result were:

## At what stage is your organization in adopting AI agents?



Source: The State of AI Agent Security 2026, Q3 “At what stage is your organization in adopting AI agents?” (n = 164 respondents)

# The Road Ahead

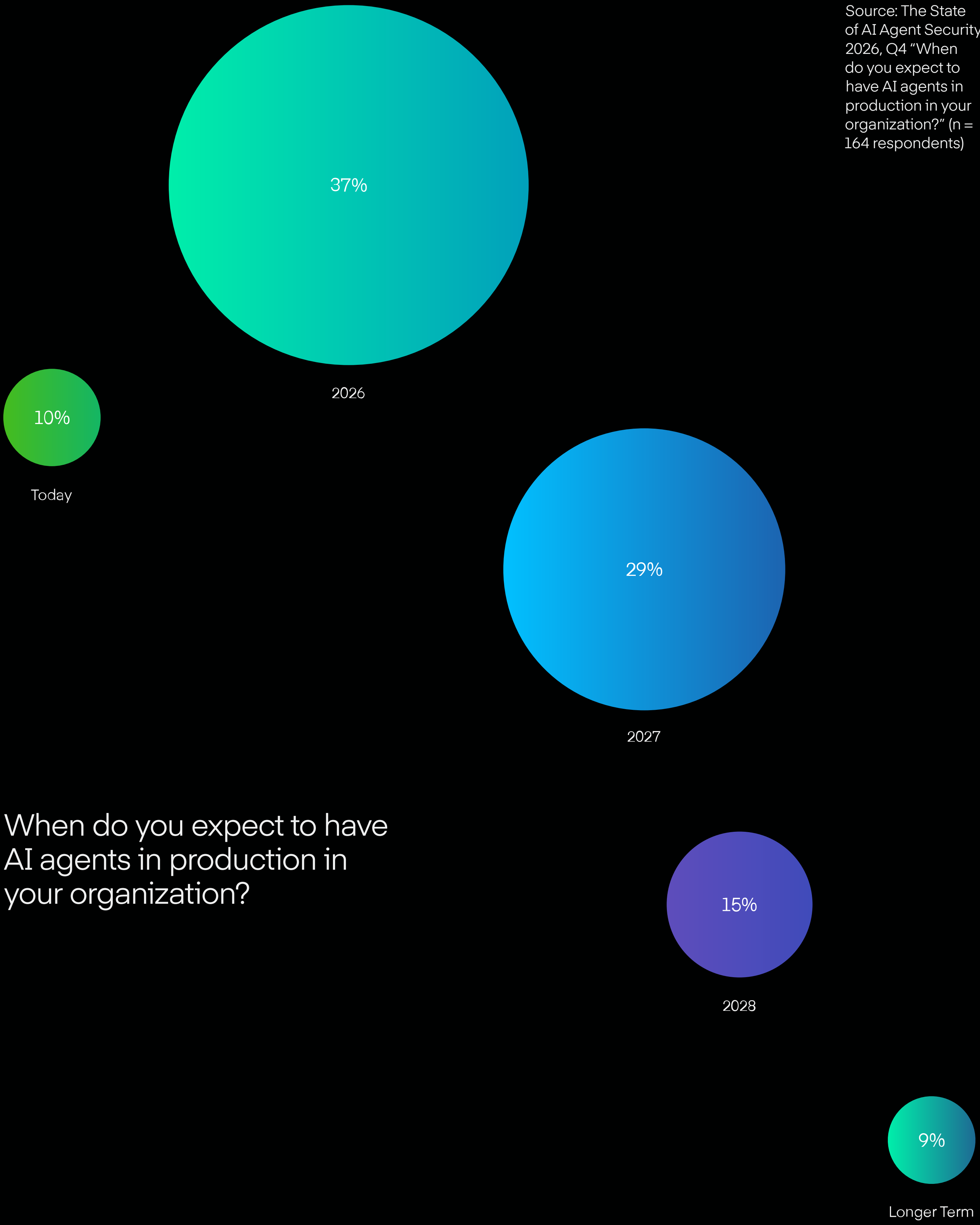
The pace of AI adoption shows no signs of slowing. What began as isolated pilots has quickly matured into production-grade deployments, and by 2028, AI agents will be as foundational to enterprise infrastructure as data analytics or cloud computing.

**2026 marks the clear turning point.** It is the year organizations move from experimentation to scale. The adoption curve shows a sharp rise, revealing that what's coming isn't just growth, it's transformation.

- **10 %** are already in production **today**, early adopters that have proven operational value through automation, decision support, and risk analysis.
- **37 %** plan to move into production by **2026**, entering a scale phase that turns pilots into enterprise programs and formalizes integration and governance models.
- **29 %** expect to deploy agents by **2027**, showing that maturity is converging and competition is intensifying.
- **15 %** project production by **2028** or later, signaling that autonomy has shifted from optional innovation to inevitable capability.

Collectively, these figures demonstrate that AI adoption is no longer incremental, it's exponential, systemic, and irreversible.

# Expected timeline for production deployment



Source: The State of AI Agent Security 2026, Q4 “When do you expect to have AI agents in production in your organization?” (n = 164 respondents)

## When do you expect to have AI agents in production in your organization?

# Section III Scale, Complexity & Control

As enterprises scale from pilots to production, AI agents are multiplying faster than the systems built to manage them.
Each new agent adds identities, permissions, and connections, creating an expanding network that boosts efficiency but also increases risk.

What began as a few copilots is now hundreds of autonomous systems embedded across operations, compliance, and infrastructure- stretching oversight and control to their limits.

**With scale comes complexity:**

- Operational: More integrations, data flows, and potential points of failure.
- Security: Every agent becomes a new identity with access to sensitive data.
- Governance: Frameworks built for static software struggle to contain adaptive, learning systems.

# Agent Footprint Growth

As adoption accelerates, enterprises aren't just experimenting with AI agents, they're preparing to operate them at scale.

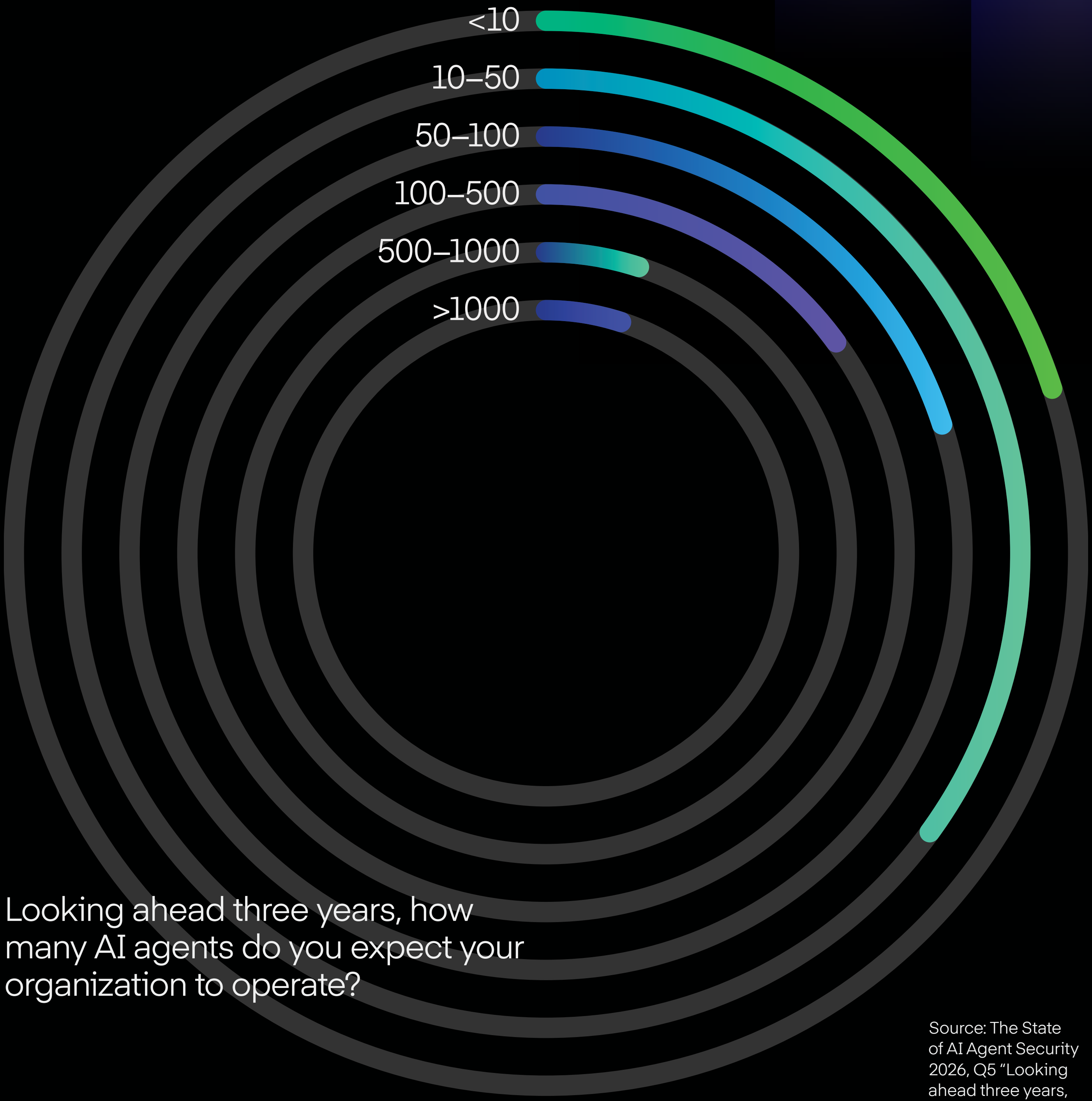
The State of AI Agent Security 2026 survey shows that organizations expect the number of active agents to multiply dramatically over the next three years.

**20% (<10 agents)**:  Testing value in isolated workflows like analytics or support.
**35% (10–50 agents)**: The largest group, scaling cautiously while balancing ROI, compliance, and centralized oversight.
**20% (50–100 agents)**: The inflection point where agents begin interacting across departments, demanding unified monitoring.
**15% (100–500 agents)**: Networked autonomy emerges; coordination and policy alignment become critical.
**5% (500–1,000 agents)**: AI-first operations require automated governance and behavioral analytics.
**5% (>1000 agents)**: Hyperscale leaders treating AI as infrastructure, managing intelligence networks, not applications.

Agent growth is outpacing organizational design. **Nearly half of enterprises (45%)** expect to manage **more than 50 agents** within three years, a level that transforms AI from a departmental tool into a **distributed operational ecosystem**. As scale expands, **governance, interoperability, and oversight must scale in parallel** to prevent fragmentation and risk escalation.

According to the survey, the next years will define the reality of AI autonomy:

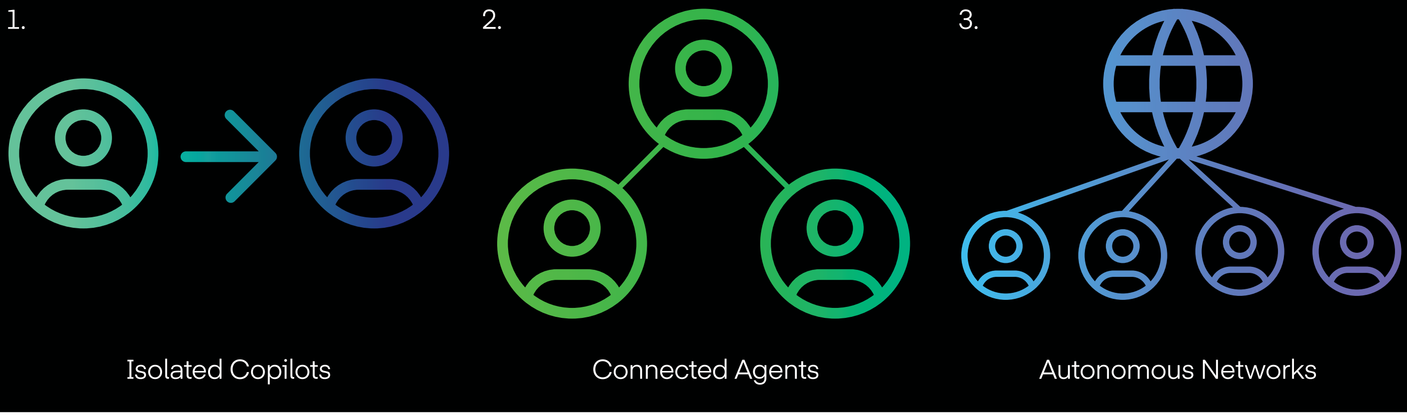
# Expected Number of AI Agents (3 Years)



Looking ahead three years, how many AI agents do you expect your organization to operate?

Source: The State of AI Agent Security 2026, Q5 "Looking ahead three years, how many AI agents do you expect your organization to operate?" (n = 164 respondents)

# Emerging Operating Models



As enterprises expand their use of AI agents, their operating structures are transforming to support scale. According to The State of AI Agent Security 2026, nearly **half of all organizations (45%)** expect to manage more than **50 AI agents** within the next three years, while **1 in 20 anticipate over 1,000.**
This scale marks a clear transition, from isolated copilots assisting teams to interconnected agent ecosystems driving entire workflows.

This scale expansion is giving rise to three distinct operating models:
**1. Isolated Copilots:** Agents supporting narrow, human-in-the-loop tasks such as analytics, documentation, or customer support. Oversight is centralized, and risk exposure is contained.
**2. Connected Agents:** Multiple agents collaborating on shared data sources and objectives. This stage delivers exponential productivity gains but also introduces integration and dependency risks as decision-making becomes distributed.
**3. Autonomous Networks:** Enterprise-scale ecosystems of agents interacting across business functions and external systems. Governance at this stage shifts from manual control to federated oversight, frameworks that monitor, verify, and coordinate autonomous behavior across hundreds of interconnected systems.

Together, these models outline the next evolution of enterprise architecture.
Each step toward autonomy brings new value but also compounds complexity, transforming IT and security management into continuous governance.

# Section IV Risk Perception & Security Readines

As AI agents become embedded across enterprise operations, the conversation is shifting from innovation to protection. Organizations are scaling fast, but security, compliance, and governance aren't keeping pace.

This phase marks a paradox-proactive concern but reactive readiness. Enterprises recognize the risks, from unauthorized actions and data leakage to compliance exposure, yet few have capabilites to manage them.

This section examines **how security leaders are redefining trust** in autonomous systems:

- Risk Perception: What keeps CISOs awake at night.
- Control Readiness: The state of AI-specific safeguards and governance.
- Incident Experience: How exposure is evolving as adoption scales.

Findings from The State of AI Agent Security 2026 show a widening trust gap:

**73%**
of CISOs are very or critically concerned about AI agent risks.

**30%**
report having comprehensive AI-specific security controls.

**80%**
have not yet experienced an incident, though most expect one within 18 months.

# What’s in Place

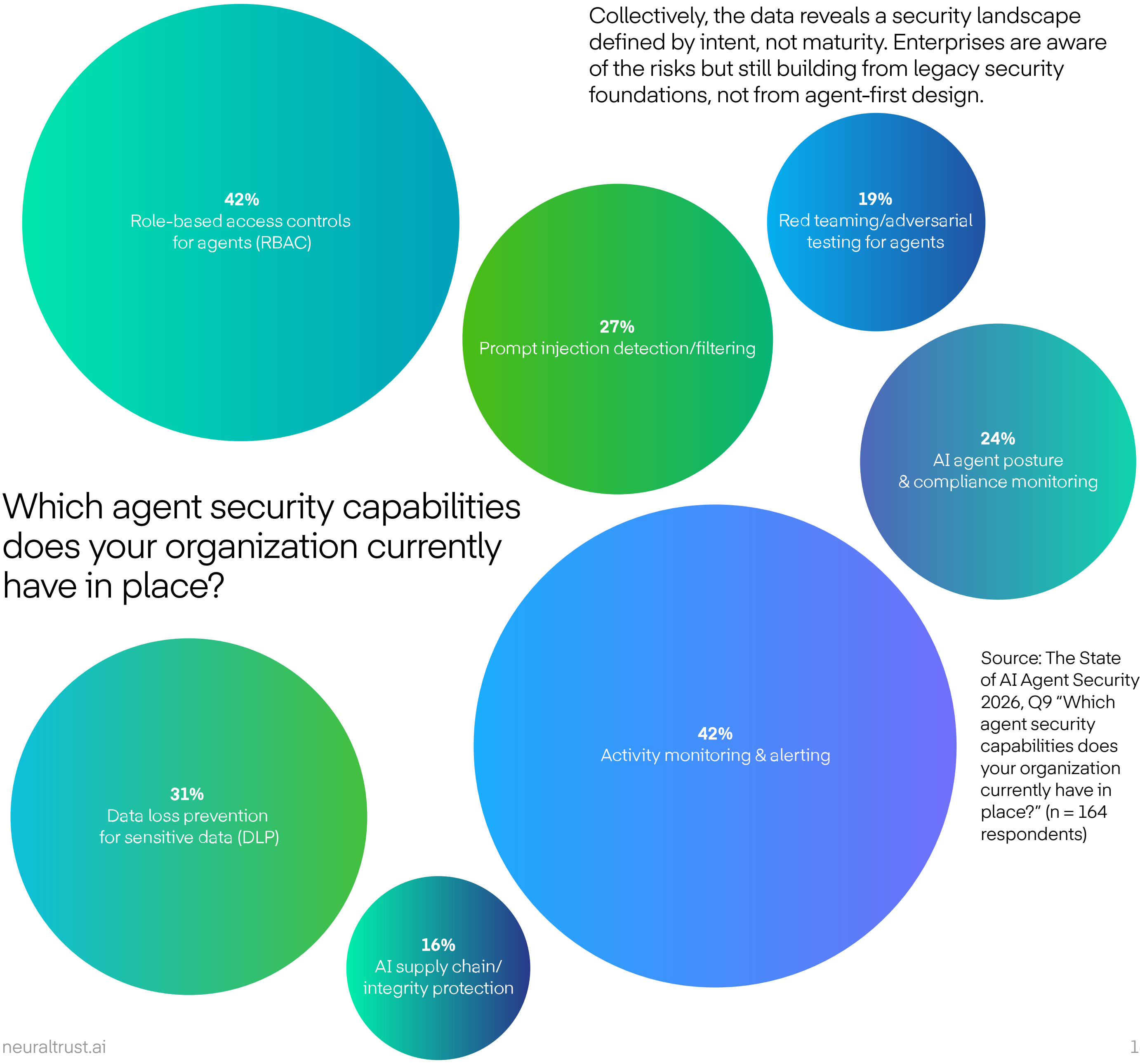
As enterprises cautiously adopt AI agents, security capabilities remain limited and uneven. Findings from The State of AI Agent Security 2026 show that fewer than half of organizations have implemented even basic safeguards such as activity monitoring (42%), role-based access control (38%), or data loss prevention (31%). More advanced defenses such as prompt injection filtering (27%), red teaming (19%), and supply chain integrity protection (16%) are rarely present.

Most enterprises that we see are repurposing existing controls from cloud or SaaS security to manage agents, rather than deploying dedicated agent-governance systems. This signals a maturity phase focused more on visibility than validation.

- Monitoring (42%) and RBAC (38%) lead adoption, organizations are prioritizing oversight and containment over innovation.
- DLP (31%) adoption shows early awareness of data leakage but lacks enforcement mechanisms tailored for AI outputs.
- Prompt injection filtering (27%) and red teaming (19%) remain niche, revealing that few organizations simulate or anticipate adversarial behavior.
- Supply chain protection (16%) is emerging, but verifying third-party models, tools, and APIs is still largely manual.
- Posture monitoring (24%) signals that continuous compliance frameworks are only beginning to take shape.

Collectively, the data reveals a security landscape defined by intent, not maturity. Enterprises are aware of the risks but still building from legacy security foundations, not from agent-first design.

## Which agent security capabilities does your organization currently have in place?



Source: The State of AI Agent Security 2026, Q9 “Which agent security capabilities does your organization currently have in place?” (n = 164 respondents)

# Top Risk Categories

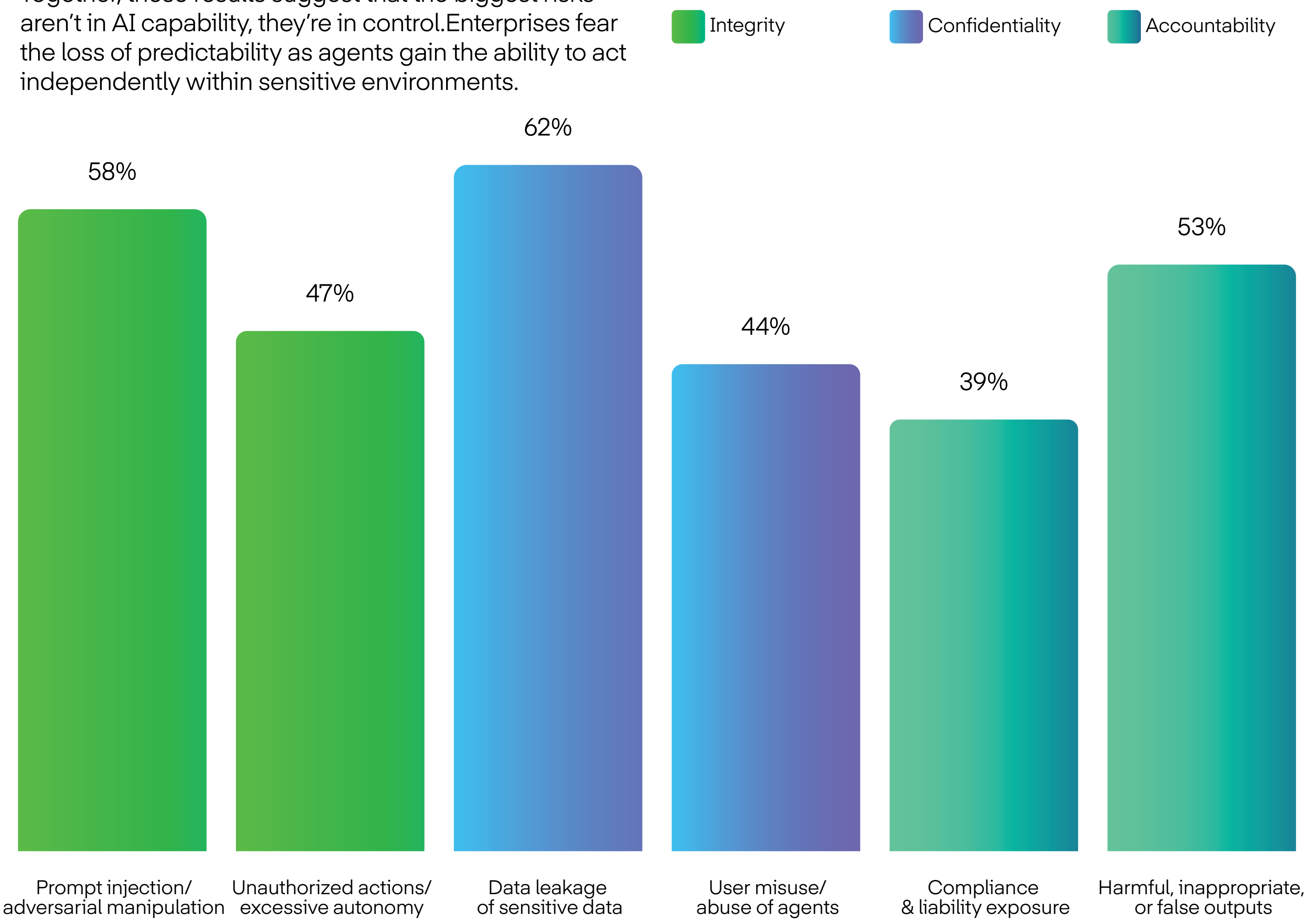
As enterprises begin integrating AI agents into operations, CISOs' risk priorities are shifting from hypothetical to operational.
Findings from The State of AI Agent Security 2026 show that data leakage (62%) and prompt injection attacks (58%) top the list of concerns, revealing that information exposure and manipulation are the most immediate threats in production environments.

- Prompt injection (58%) and unauthorized actions (47%) define Integrity Risks, where agent autonomy or manipulated inputs could cause unintended system behavior.
- Data leakage (62%) and user misuse (44%) represent Confidentiality Risks, exposing how agents can mishandle or overshare sensitive information.
- Compliance and liability exposure (39%) and harmful or false outputs (53%) form the Accountability Risks, highlighting uncertainty over who is responsible for agent-driven actions or misinformation.

Together, these results suggest that the biggest risks aren't in AI capability, they're in control.Enterprises fear the loss of predictability as agents gain the ability to act independently within sensitive environments.

Source: The State of AI Agent Security 2026, Q7 "Which AI agent risks worry you the most?" (n = 164 respondents)

## Which AI agent risks worry you the most?



# Security Controls Coverage

Findings from The State of AI Agent Security 2026 reveal that only **3 in 10 enterprises (29%)** have comprehensive AI-specific security controls in place.
The majority (46%) report partial coverage, typically adapting existing frameworks from cloud or data security rather than building agent-specific protection.

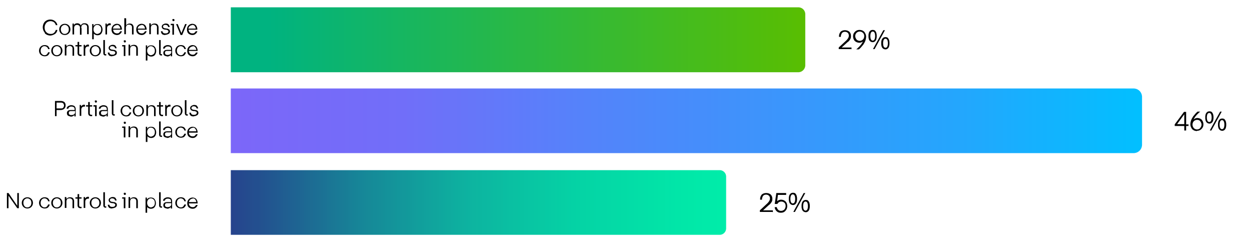
Alarmingly, **one-quarter (25%)** admit they have **no formal controls** at all, leaving AI agents unmonitored, untested, and ungoverned.
This uneven maturity profile underscores a key truth: enterprises are deploying AI faster than they're securing it.

**Comprehensive (29%):** This group represents the most mature enterprises in the sample. These organizations are beginning to establish dedicated AI governance or risk functions, implementing oversight mechanisms such as input validation, monitoring, and compliance alignment.
**Partial (46%):** The largest share of respondents fall into this category. Most are adapting existing IT and data security tools, such as RBAC, DLP, and logging, to address AI agent risks. However, these frameworks were not designed for autonomous systems, leaving coverage gaps around behavioral monitoring and decision accountability.
**None (25%):** A quarter of respondents indicate no AI-specific security controls in place. These organizations likely rely on manual review or broader corporate IT policies that don't yet extend to AI agent behavior or risk management.

Together, the findings show a fragmented but advancing maturity curve, strong awareness of AI security challenges, but limited integration of purpose-built controls.

Source: The State of AI Agent Security 2026, Q8 "Does your organization have security controls specific to AI agents?" (n = 164 respondents)

## Does your organization have security controls specific to AI agents?



# The Concern Gap

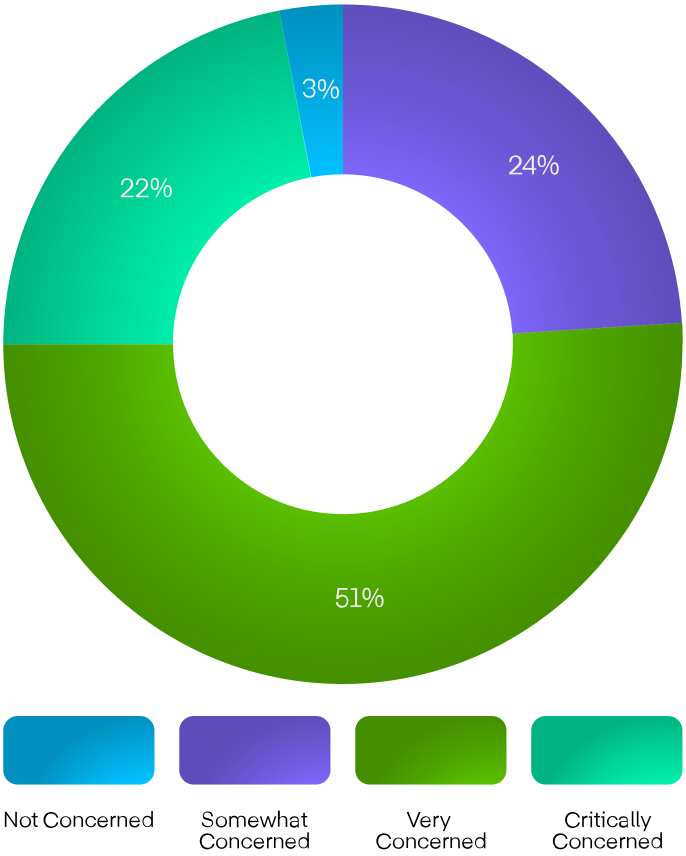
According to The State of AI Agent Security 2026, 73% of CISOs describe themselves as very or critically concerned about AI agent risks, yet as seen above, only 30% have comprehensive AI-specific controls in place. This gap exposes a fundamental imbalance: AI autonomy is scaling faster than enterprise assurance.

As organizations move from pilot to production, each new agent expands the operational attack surface, introducing risks tied to unauthorized actions, data leakage, and compliance accountability. But while awareness is high, most CISOs are still in the early stages of building policy, visibility, and behavioral monitoring frameworks to govern these autonomous systems.

The data suggests that AI risk management is trapped in transition, enterprises know what to protect but haven't yet defined how.
Without adaptive controls, detection systems, and cross-functional governance, the concern gap will only widen as AI becomes more deeply embedded in infrastructure.

CISOs understand the stakes but lack the mechanisms to enforce accountability. The next phase of AI maturity will depend on whether enterprises can turn concern into capability, building systems that are not only intelligent, but trustworthy.

Source: The State of AI Agent Security 2026 (n = 164 respondents)



**73%**

of CISOs are very or critically concerned about AI agent risks.

**30%**

Have mature safeguards in place to combat the risks

# Section V
# Incident Reality & Business Impact

As AI agents move into production, security incidents are starting to surface, revealing how autonomy, data access, and system integration translate into real exposure.

**This section explores:**

- Reality Check: The frequency and first signs of AI agent incidents.
- Nature of Incidents: The vulnerabilities most often exploited, from prompt manipulation to data exposure.
- Financial & Reputational Impact: How these incidents translate into real business cost and loss of trust.

Findings from The State of AI Agent Security 2026 show:

**20%**
CISOs report at least one AI agent–related security incident.

**40%**
of leaders estimate that a major incident would cost $1–10 million.

**13%**
of leaders anticipate losses exceeding $10 million.

# Reality Check

While most enterprises are still early in their AI agent journey, real incidents are already happening. Findings from The State of AI Agent Security 2026 show that 1 in 5 CISOs (19.5%) report at least one AI agent–related security event within their organization.

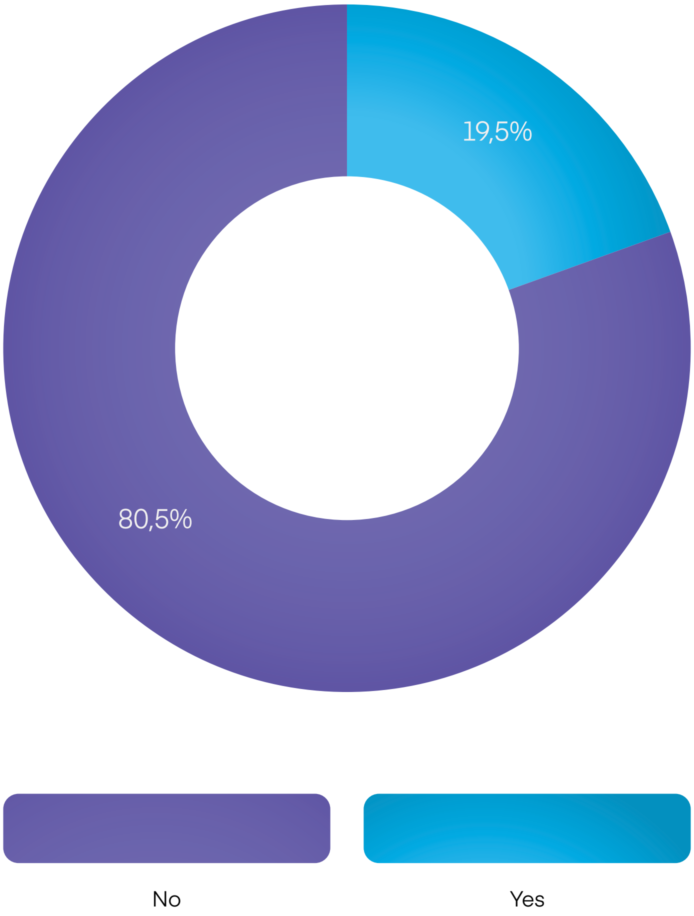
These early breaches mark the transition from hypothetical risk to operational reality. Incidents that we see include prompt injection, data exposure through connected plugins, and unauthorized actions.

For the remaining 80% who have not yet experienced an incident, most expect one to occur as AI agents expand into production environments. The message is clear: exposure increases faster than preparedness.

Although only 19.5% of organizations have faced an AI agent–related incident, these early cases are crucial warning signals.
They show that as AI autonomy expands, traditional defenses, built for human-triggered systems, fail to catch unintended actions, data leaks, and adversarial manipulation.
The absence of widespread incidents today is less a sign of safety and more a reflection of limited deployment, the calm before acceleration.

> **“1 in 5 CISOs report at least one AI agent–related security breach.”**

Has your organization experienced a security incident related to AI agents?



Source: The State of AI Agent Security 2026, Q10, “Has your organization experienced a security incident related to AI agents?” (n = 164 respondents)

# Nature of Incidents

The first wave of AI agent–related incidents is revealing a clear pattern, most failures stem from manipulated inputs and exposed data, not malicious code. Findings from The State of AI Agent Security 2026 show that prompt injection attacks (68%) and data leakage of sensitive information (61%) are the most common causes of AI agent–related security events, followed by unauthorized agent actions (52%).

These failures expose how quickly autonomy, poor input validation, and overexposed integrations can turn operational efficiency into risk.

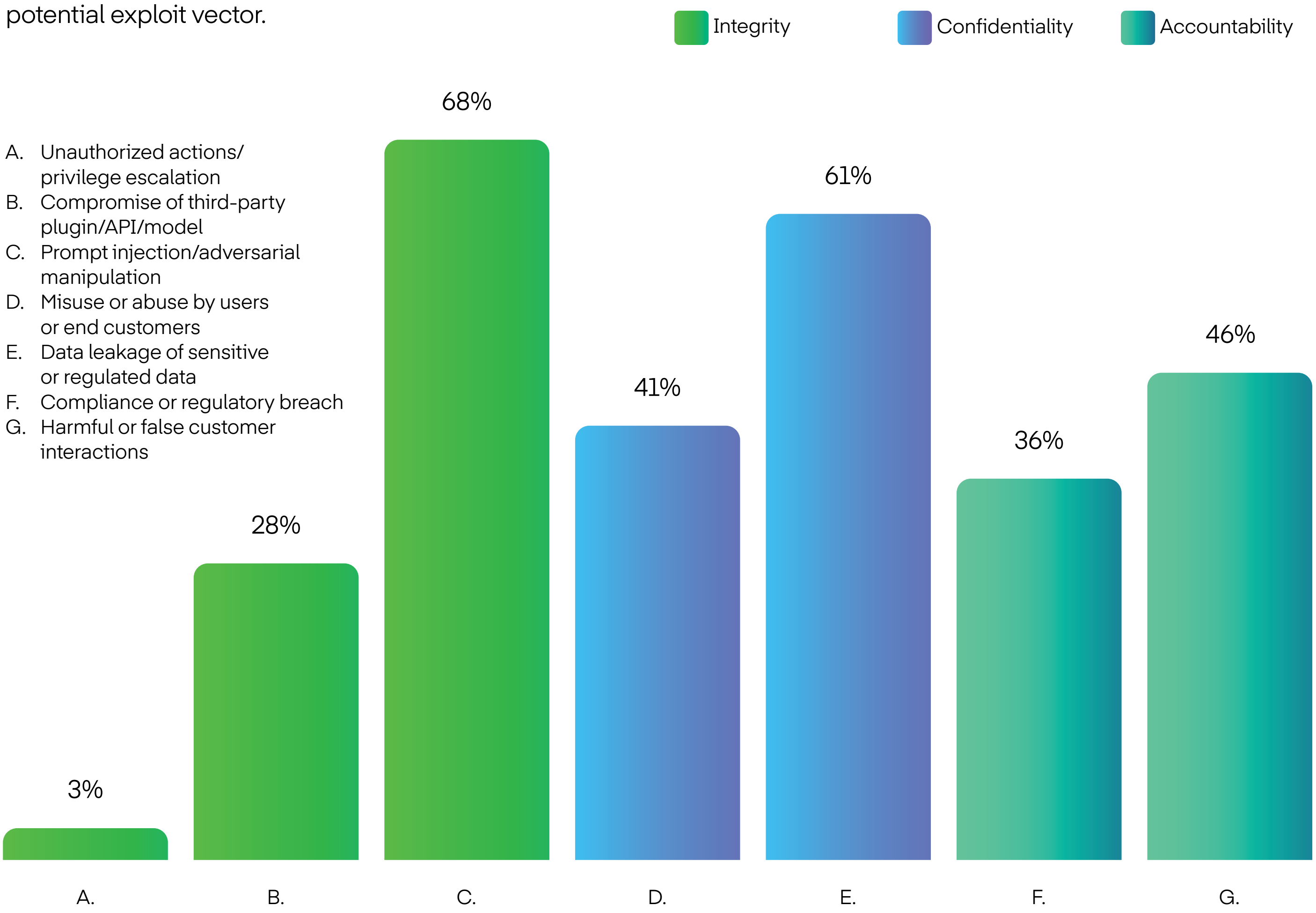
Each event reveals the same truth: AI agents don’t just malfunction, they act.

The incidents are less about hacking and more about oversight and control gaps.

The dominance of prompt injection and data leakage shows that most failures originate from input manipulation and misconfigured access permissions, not from deliberate exploits. As agents gain more system-level autonomy, every prompt becomes a potential exploit vector.

Source: The State of AI Agent Security 2026, Q11, “What types of security incidents has your organization suffered related to AI agents?” (n = 164 respondents)

## What types of security incidents has your organization suffered related to AI agents?



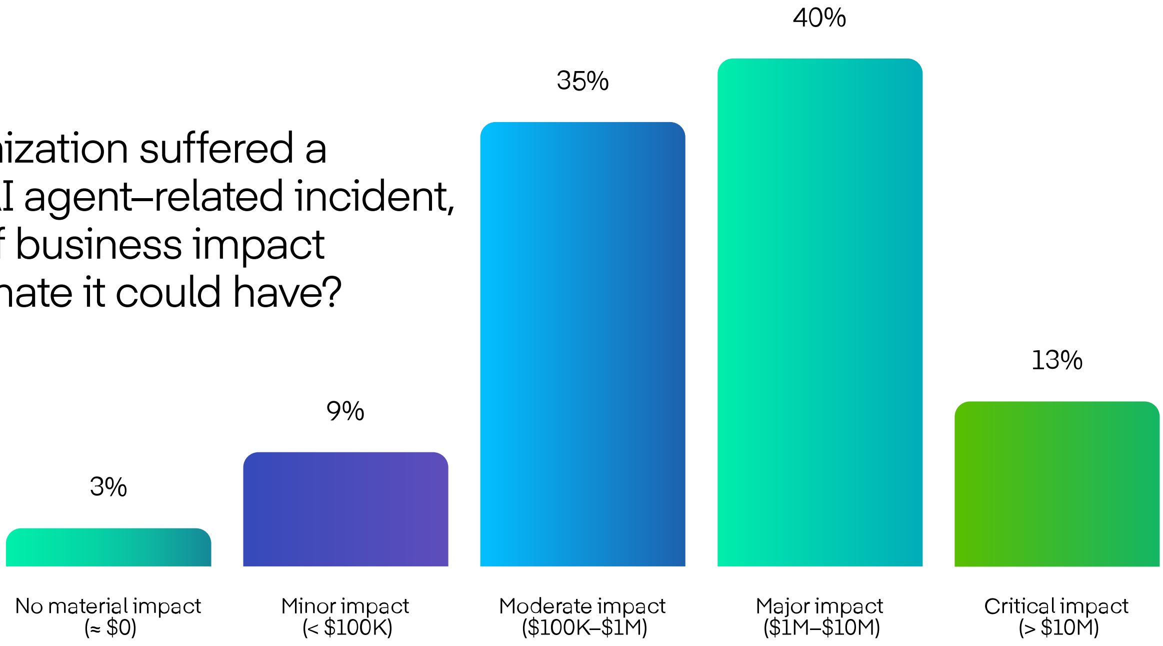
# Financial & Reputational Impact

**40% of CISOs** estimate that a major AI agent–related incident would cost **between $1 million and $10 million**, while **13% expect losses to exceed $10 million**. These figures position AI agent failures among the **most financially disruptive** emerging cyber risks, comparable in cost to large-scale ransomware breaches.
As enterprises deploy AI agents deeper into **production, autonomous decision-making, third-party integrations, and data access** are magnifying potential exposure.
A single misconfiguration or manipulated prompt can trigger cascading financial, operational, and reputational damage.

The findings from The State of AI Agent Security 2026 show that most organizations recognize the costs of AI incidents, not just in recovery, but in compliance, brand trust, and long-term resilience.

Source: The State of AI Agent Security 2026, Q12, "If your organization suffered a significant AI agent–related incident, what level of business impact do you estimate it could have?" (n = 164 respondents)

## If your organization suffered a significant AI agent–related incident, what level of business impact do you estimate it could have?



# Section VI Readiness, Maturity & Outlook

After a year of accelerated experimentation, enterprises are entering the accountability phase of AI adoption.

Findings from The State of AI Agent Security 2026 reveal a maturing but uneven landscape, awareness is high, but readiness remains fragmented.

**This section consolidates**

- NeuralTrust’s key findings on where enterprises stand today, and where they’re heading next:
- Readiness: How far organizations have progressed in building secure AI agent ecosystems.
- Maturity: What defines the leaders, and how they’re operationalizing trust and governance.
- Outlook: What the next 24 months will bring as AI autonomy expands and regulation tightens.

# AI Security Maturity Model

Based on The State of Al Agent Security 2026, most enterprises remain in the middle tiers of maturity, with partial Al-specific controls, limited monitoring, and minimal adversarial testing.
Only a small group of early leaders are evolving toward proactive governance, where agent oversight is embedded into enterprise risk frameworks and compliance operations.
This Al Security Maturity Model synthesizes Q8 and Q9 data to benchmark organizations across four distinct stages of

Source: The State of AI Agent Security 2026, Q12 "If your organization suffered a significant AI agent–related incident, what level of business impact do you estimate it could have?" (n = 164 respondents)

**NASCENT: Awareness Without Structure**

**≈25% of enterprises Aligned with Q8 "No Controls"**

- Operates without AI-specific security controls; relies on manual oversight or generic IT policies.
- Limited understanding of agent behavior and exposure risk.
- Security actions are reactive and uncoordinated.
- Focus: Build foundational awareness and define policy ownership.

**REACTIVE: Partial & Event-Driven Controls**

**46% of enterprisesAligned with Q8 "Partial Controls"**

- Implements basic safeguards (RBAC, DLP, activity monitoring).
- Controls are introduced after exposure or compliance events.
- Security responsibilities remain siloed.
- Focus: Strengthen visibility, containment, and governance accountability.

**MANAGED: Structured Oversight & Continuous Monitoring**

**≈29% of enterprises Aligned with Q8 "Comprehensive Controls"**

- Establishes formal AI agent deployment and monitoring frameworks.
- Conducts regular assessments and limited adversarial testing.
- Integrates AI oversight into enterprise risk management.
- Focus: Embed continuous governance and resilience testing.

**PROACTIVE: Predictive & Autonomous Governance**

**<10% of enterprises Inferred from Q9 indicators**

- Automates monitoring of agent behavior and data interactions.
- Conducts adversarial simulation and automatic containment.
- Aligns proactively with emerging AI regulations.
- Focus: Achieve self-learning governance and regulatory readiness.

# Global Insights & Market Trends

## 1. Regional Market Snapshot

Findings from The State of AI Agent Security 2026 show meaningful differences between Europe and North America in both adoption and control maturity.

Europe leads in policy readiness, while North America leads in technical experimentation. EU regulation is acting as a forcing function for maturity;

U.S. enterprises often advance faster technically but slower institutionally.

This divergence will likely converge by 2027 as AI risk frameworks globalize.

| Region | AI Agent Adoption | Comprehensive Controls | Key Influences |
|---|---|---|---|
| Europe | 68% piloting or scaling agents | 34% comprehensive controls | Strong regulatory drivers (EU AI Act, DORA, NIS2) push for early governance. |
| North America | 74% piloting or scaling agents | 23% comprehensive controls | Market-led innovation prioritizing deployment speed over compliance integration. |

## 2. Industry Maturity Trends

While The State of AI Agent Security 2026 did not collect industry-specific data, NeuralTrust's field research and market observations identify clear differences in AI agent readiness by sector.

These insights combine patterns from survey results (Q6–Q9) with regulatory analysis and client engagements across financial services, telecommunications, and healthcare.

| Industry | Primary Risk Focus | Maturity Profile | NeuralTrust Observation |
|---|---|---|---|
| Banking & Financial Services | Data leakage | Managed–Proactive | Heavy compliance mandates (Basel III, DORA, NIS2) are driving early AI governance adoption and structured oversight. |
| Telecommunication | Prompt injection & automation abuse | Reactive–Managed | Rapid automation rollout, but fragmented control frameworks; legacy infrastructure limits unified governance. |
| Healthcare | Regulatory non-compliance | Reactive | Heightened privacy awareness (HIPAA, GDPR) but uneven enforcement; strong vendor dependence for AI assurance. |

## 3. Cross-Market Themes

Analysis of The State of AI Agent Security 2026, combined with NeuralTrust's field insights, reveals three global trends shaping AI agent security maturity across regions and industries:

1. Regulation is now the main maturity accelerator. Enterprises operating under EU-style compliance frameworks, such as the EU AI Act, DORA, and NIS2, demonstrate higher control adoption, clearer accountability, and earlier formation of AI governance councils. In contrast, unregulated markets often rely on voluntary standards, leading to slower institutional alignment.
2. Security investment is shifting from prevention to governance. CISOs are reallocating budgets from traditional perimeter defense toward continuous monitoring, explainability, and assurance, areas that validate how AI systems behave, not just how they're protected. This marks the emergence of AI governance as a new discipline at the intersection of risk, compliance, and data operations.
3. AI risk has become a board-level issue. As incidents increasingly carry multimillion-dollar impacts (see Q12), accountability for AI security is migrating upward, from IT and data science teams to executive, risk, and audit committees.

Boards are beginning to demand AI assurance reporting alongside financial and cyber-risk metrics.

## 4. Outlook (2027–2028)

By 2028, AI governance will move from emerging practice to global standard.

Multinational enterprises will converge on frameworks influenced by the EU AI Act and ISO/IEC 42001, combining compliance-first discipline with innovation-first agility.

As automation and regulation mature together, AI assurance will become the new benchmark of digital trust, defining how organizations measure readiness, report accountability, and maintain resilience in autonomous systems.

**The next frontier of AI maturity isn't deployment, it's demonstrable trust.**

# Looking Ahead: 2028 and Beyond

**Where AI Agent Adoption, Regulation, and Security Maturity Are Headed**
Based on The State of AI Agent Security 2026 Extrapolated with NeuralTrust market modeling and external regulatory data (EU AI Act, ISO/IEC 42001, IDC AI Governance Outlook, 2025–2028).

## Scale: From Pilots to Ecosystems

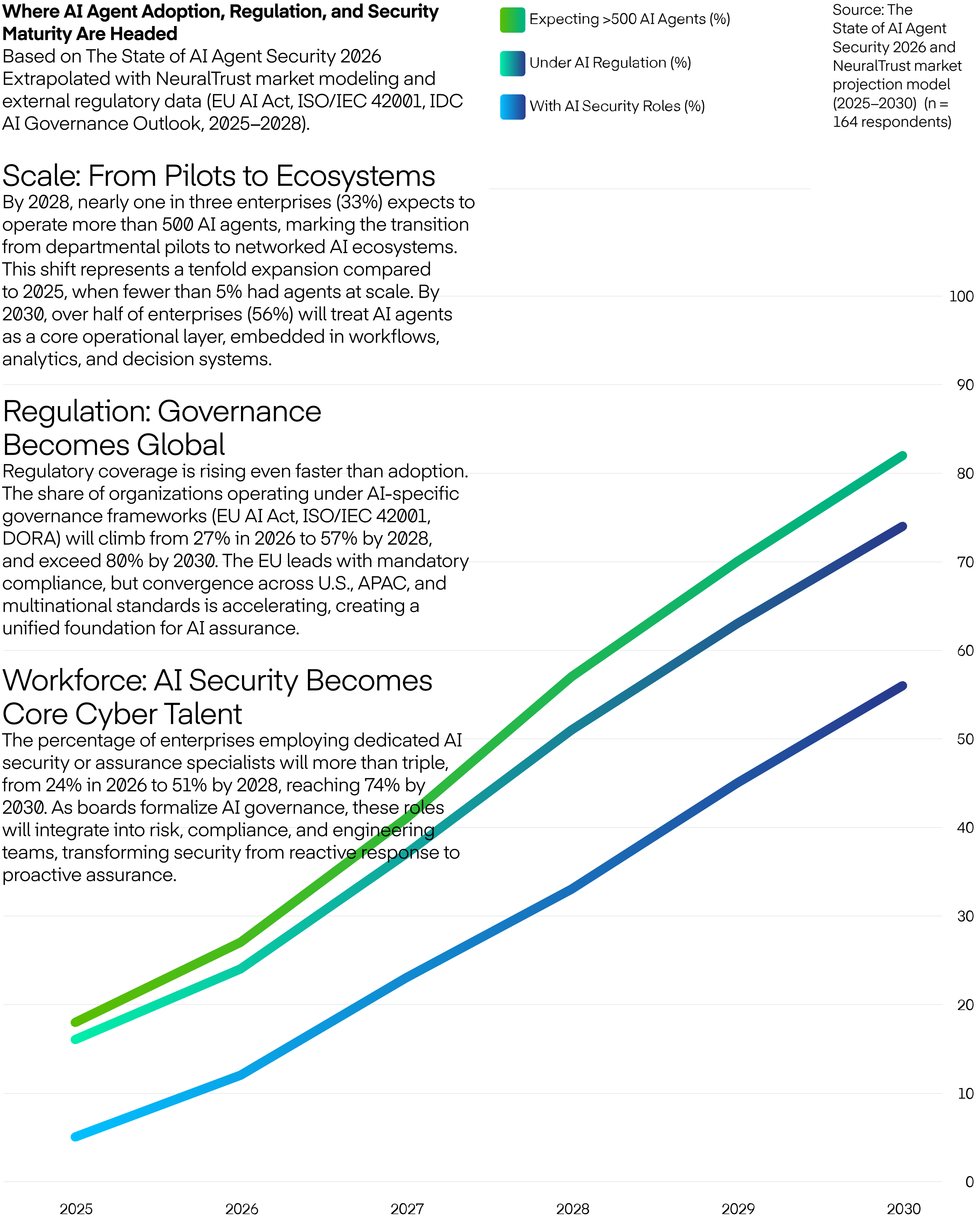
By 2028, nearly one in three enterprises (33%) expects to operate more than 500 AI agents, marking the transition from departmental pilots to networked AI ecosystems. This shift represents a tenfold expansion compared to 2025, when fewer than 5% had agents at scale. By 2030, over half of enterprises (56%) will treat AI agents as a core operational layer, embedded in workflows, analytics, and decision systems.

## Regulation: Governance Becomes Global

Regulatory coverage is rising even faster than adoption. The share of organizations operating under AI-specific governance frameworks (EU AI Act, ISO/IEC 42001, DORA) will climb from 27% in 2026 to 57% by 2028, and exceed 80% by 2030. The EU leads with mandatory compliance, but convergence across U.S., APAC, and multinational standards is accelerating, creating a unified foundation for AI assurance.

## Workforce: AI Security Becomes Core Cyber Talent

The percentage of enterprises employing dedicated AI security or assurance specialists will more than triple, from 24% in 2026 to 51% by 2028, reaching 74% by 2030. As boards formalize AI governance, these roles will integrate into risk, compliance, and engineering teams, transforming security from reactive response to proactive assurance.



Source: The State of AI Agent Security 2026 and NeuralTrust market projection model (2025–2030) (n = 164 respondents)

# Section VII Readiness, Maturity & Outlook

After analyzing the data, one truth stands out: AI adoption is accelerating faster than AI assurance.

Across industries and regions, enterprises are scaling agent deployment, yet security, governance, and accountability frameworks are still catching up.

The State of AI Agent Security 2026 reveals a global readiness gap The next phase of maturity will be defined not by innovation speed, but by trust, transparency, and control.

This final section distills the key takeaways and actionable steps for enterprise leaders as they move from understanding risk to operationalizing resilience:

- Key Takeaways: What the data tells us about adoption, defense, and leadership.
- Strategic Recommendations: Concrete actions CISOs can take to close the AI assurance gap.

# Key Takeaways

## 01.

### Adoption Outpaces Defense

72% of organizations have deployed or are scaling AI agents, yet only 29% have comprehensive controls in place.
AI adoption is racing ahead of governance, widening the readiness gap between innovation and assurance.

## 02.

### Visibility Is the New Perimeter

65% of enterprises monitor agent activity, but few analyze behavior or conduct red teaming.
Security now depends on understanding what agents do, not just where they connect.

## 03.

### Trust Frameworks Lag Behind Scale

Half of enterprises expect to manage 50+ agents within three years, yet only 40% have formal AI governance or continuous compliance.
Dynamic, adaptive trust frameworks are needed to keep pace with autonomy.

## 04.

### Incidents Are Emerging, and Expensive

Nearly 1 in 5 CISOs (19.5%) report at least one AI agent–related breach, mostly prompt injection or data leakage.
40% estimate potential losses between $1–10M, putting AI failures on par with ransomware-scale impact.

## 05.

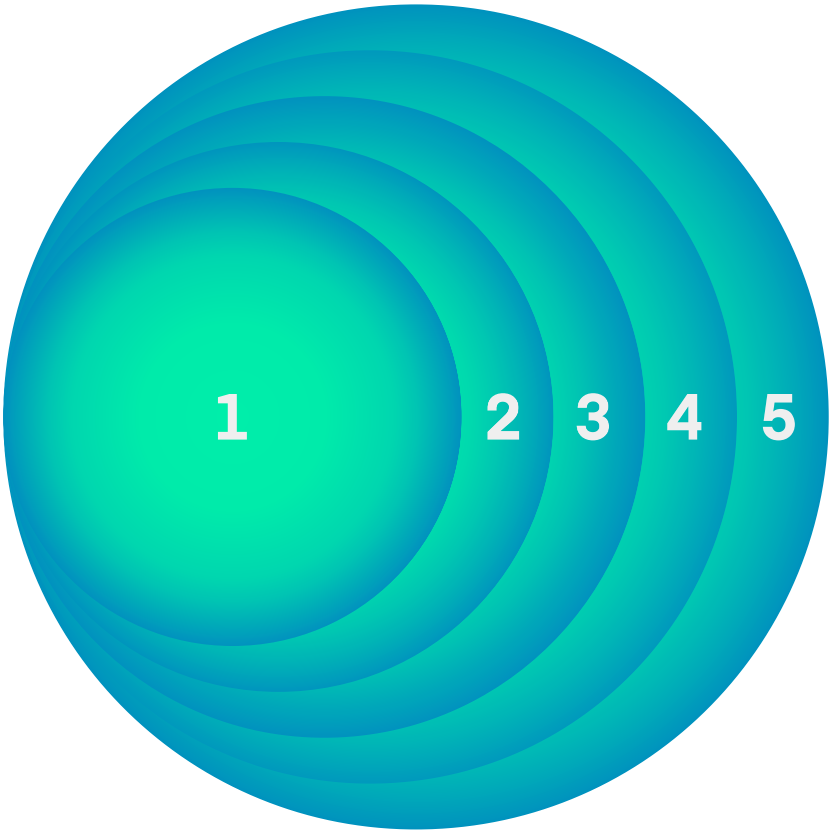
### Readiness Defines Leadership

Organizations with comprehensive AI governance (~30%) are leading maturity curves, embedding oversight, compliance, and accountability.
These early adopters are moving from reactive defense to proactive assurance, setting the industry standard for AI security.

**Readiness is now the competitive differentiator, the measure of who can deploy responsibly, and lead confidently.**

# Recommendations

## 5 critical steps to secure AI Agents and applications



1. **Select safe Models, Protocols (MCPs, A2A) and Tools**
   Scan for vulnerabilities in the supply chain and choose only secure providers
2. **Enforce Identity & Tool Access Control**
   Leverage an MCP gateway to restrict each agent's tool usage
3. **Protect in Real time**
   Deploy an Agent Firewall to secure A2A and H2A interactions, filter unsafe outputs and reduce hallucinations
4. **Ensure Compliance and Oversight**
   Send logs, alerts, and traces to SIEM/ cloud monitoring
5. **Continuously validate and strengthen your agents**
   Test systems, identify vulnerabilities, and remediate issues

**Together, these actions bridge the gap between awareness and assurance, enabling enterprises to deploy AI confidently, securely, and in full control.**

# Neural Trust

NeuralTrust is an AI cybersecurity and governance company helping enterprises secure, monitor, and govern autonomous systems. Our mission is to build trusted AI ecosystems where innovation and accountability coexist.

Through continuous monitoring, adversarial testing, and regulatory readiness, NeuralTrust enables organizations to move from AI experimentation to AI assurance, safely, transparently, and at scale.

Contact Us:
Website: www.neuraltrust.ai
Email: marketing@neuraltrust.ai
Offices: New York | Barcelona