



# From Deepfakes to Double Agents

The New Face of AI-Driven Cyber Risk



Increased investment in deepfake defense is beginning to pay off, but the rapid rise of agentic AI is opening organizations up to a whole new world of risk.

CYBERSECURITY  
THREAT REPORT 2026

# Executive Summary



In our third annual Cybersecurity Threat Report, produced with market research firm Censuswide, the data tells a story in two parts: **organizations followed through on their promise to fight back against deepfakes, but while they did, a new AI-driven threat emerged in the enterprise.**

Last year, most of the IT and cybersecurity professionals we surveyed told us their organizations had yet to invest in deepfake defense, but planned to soon. This year, they delivered. The share of organizations already investing in deepfake defense rose from 37% to 48%. And **among organizations that lost money to deepfake-related incidents<sup>1</sup>, losses were less severe**, with the estimated average falling from roughly \$280,000 to roughly \$225,000.

But while security teams were catching up on deepfakes, AI took on a new role inside their organizations. **96% of respondents say their organizations now use AI agents or AI tools that**

37% **TO**  
48%

can take actions on behalf of employees<sup>2</sup>, and **77% say one of those tools has already acted on instructions or content that turned out to be malicious, misleading, or unauthorized.**

The most common way in? The inbox.

Meanwhile, the teams tasked with security are largely absent from decisions about these tools. **IT is the primary approver of AI agents at 78% of organizations, while security is at just 6%.**

And just as we saw with deepfakes last year, confidence remains high in spite of widespread incidents. The preparedness paradox is back, this time the disconnect is around AI agents.

<sup>1</sup> Deepfake-related incidents encompass attacks using synthetic or manipulated media, from basic image alterations and doctored emails to sophisticated real-time voice and video impersonations.

<sup>2</sup> Throughout this report, "AI agents" refers to AI agents or AI tools that can take actions on behalf of employees, such as reading and responding to messages, accessing files and systems, or triggering workflows.

In this report, we share the data behind both stories and offer recommendations on what organizations can do to stay ahead of today's changing threat landscape. The following five key takeaways offer a snapshot of the report's findings and what they mean for security teams moving forward:

#### KEY TAKEAWAYS

# Orgs Kept Their Promises on Deepfake Defense:

# Investment Is Up



# and Losses Are Less Severe



Last year's respondents told us deepfake defense investment was on the way. This year's data shows they meant it. And while the survey can't prove for certain that increased investment caused lower losses, the two trends clearly moved in the same direction. In our survey, we found that:

48%

**of organizations are now actively investing in deepfake defense, up from 37% in 2025 (an 11-point increase).**

That growth came directly from organizations that had planned to invest. The share planning to invest "within 5 months" fell from 32% to 27%, "within 6–11 months" from 22% to 17%, and "within 12–18 months" from 7% to 5%.

Among organizations reporting a financial loss, **the share losing \$100,000 or more fell from 63% to 49%.** Losses of \$500,000 or more fell from 20% to 17%, and losses of \$1 million or more fell from 6% to 3%.

**The estimated average loss per affected organization fell about 20%,** from roughly \$280,000 to roughly \$225,000.

Repeat attacks are down: **organizations hit three or more times fell from 40% to 35%.** However, the share experiencing at least one incident held steady (85% in 2025 vs. 84% in 2026).

# AI Agents Are Now Commonplace in the Enterprise

Agentic AI has moved well past the exploratory phase. The technology is already being used at nearly every organization we surveyed, and it has access to some of their most sensitive and consequential systems.



# 96%

of organizations use AI agents or tools that can take actions on behalf of employees, and 52% say that use is widespread across the organization.

AGENTS HAVE BROAD ACCESS,  
MOST COMMONLY TO

# 67%

TICKETING AND  
HELPDESK SYSTEMS

# 56%

CLOUD STORAGE

# 49%

EMAIL AND MESSAGING PLATFORMS  
(THE THIRD MOST COMMON)

# 49%

INTERNAL FILE REPOSITORIES

# 47%

SECURITY TOOLS AND LOGS

Even finance, billing, and payment systems are within reach of these tools at 30% of organizations.

Only 19% limit agents to the public web, meaning **roughly 4 in 5 organizations give AI agents access to internal systems.**

# AI Agents Are Already Being Exploited, and Email Is the Top Entry Point

Attackers have noticed the new access AI agents enjoy, and they're using some of the oldest tricks (and the oldest entry points) in the book to exploit it.

**77% say an AI agent or tool acted on instructions or content that turned out to be malicious, misleading, or unauthorized in the past 12 months.** 57% of all organizations have a confirmed case.

Among organizations with incidents, the two most common incident types were an AI tool summarizing or acting on a malicious email or message (**42%**) and an attacker using an AI tool in a phishing or social engineering attempt (**41%**).

Respondents agree on where the risk lies: **acting on a malicious email, prompt, file, or link is the #1 perceived way** an AI agent could cause a security incident (23%).



77%

**say an AI agent or tool acted on instructions or content that turned out to be malicious, misleading, or unauthorized in the past 12 months.**

# The Security Disconnect: Cybersecurity Teams Have Little Say Over Agentic AI

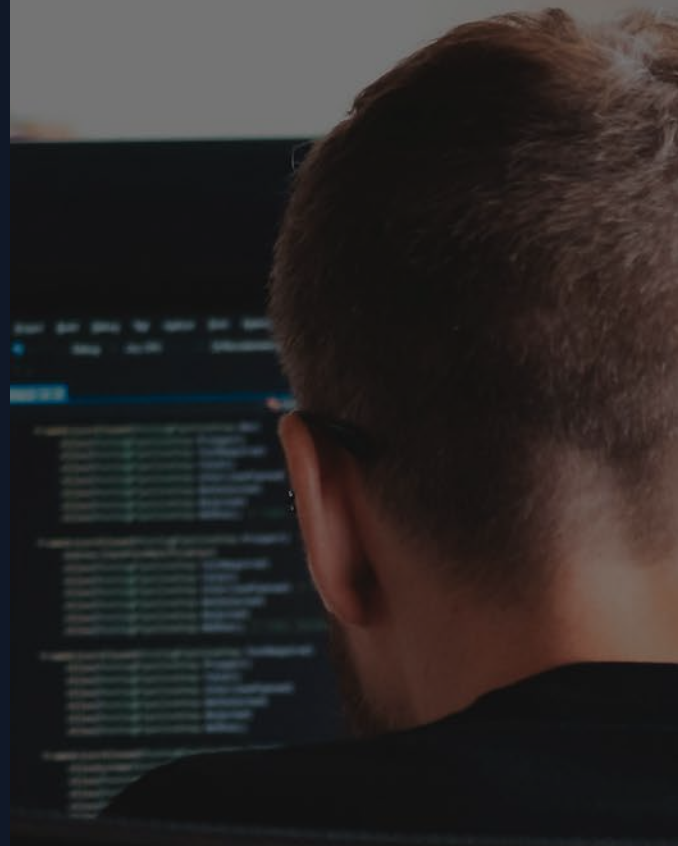
The teams responsible for defending their organizations are largely missing from decision-making about which AI agents get deployed and what they are able to access and do.

**IT is the primary approver of AI agents, plugins, and connectors at 78% of organizations. Security plays this role at just 6%.**

Logically, IT is also seen as primarily responsible if an agent causes harm (64%, vs. 9% for security), and respondents name IT as the department creating the greatest amount of AI-agent-related risk (66%) at their organizations.

# 39%

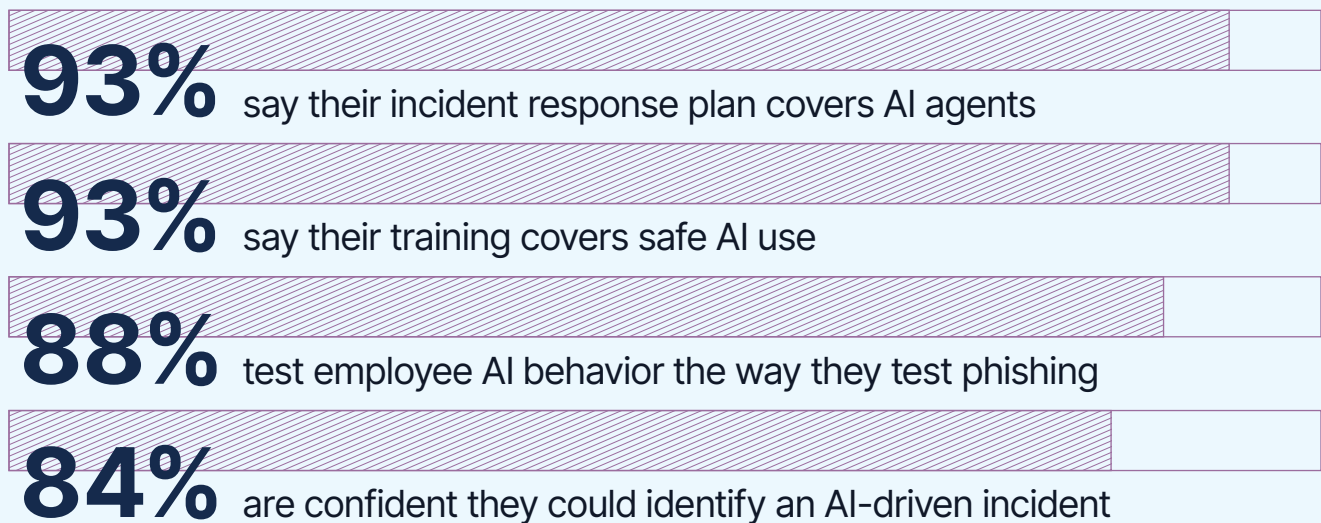
**of organizations that have experienced agentic-related incidents say an employee gave an AI tool access that security would not have approved.**



# The Return of the Preparedness Paradox

Last year, we found near-universal confidence in deepfake defenses despite widespread deepfake-driven compromise and massive associated financial losses. This year, the same gap has opened up around agentic AI. While confidence abounds, the data suggests otherwise.

On paper, organizations look ready for the age of agentic AI:



Yet 77% have already had an AI agent act on malicious or unauthorized content, and **only 52% have formal policies, controls, and incident response processes in place.**

Deepfakes still matter: concern remains high at 96%, and deepfake detection is the #1 planned defensive investment (39%). But **the share calling deepfake defense a "top priority" slipped from 71% to 60%** as agentic AI competes for organizations' attention.

## SECTION 1

# Deepfake Defense: Promises Kept, Losses Less Severe

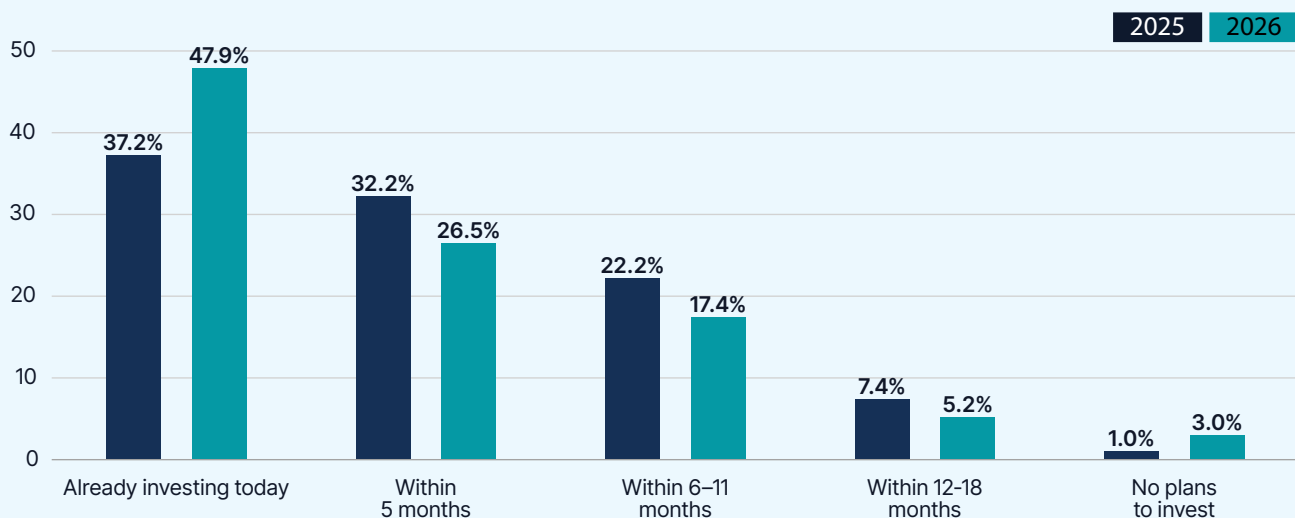
In last year's report, we found organizations in an uncomfortable position. Deepfake-related incidents had become a fact of life, with 85% of respondents reporting having experienced at least one such event in the past 12 months and the average financially impacted organization losing more than \$280,000.

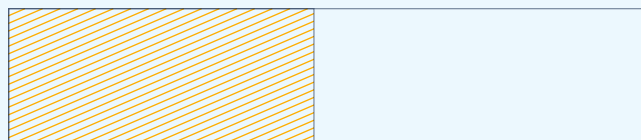
Yet 63% of organizations had not invested a single dollar in deepfake defense. The encouraging part, however, was intent: only 1% said they had no plans to invest in deepfake-defense, and most expected to do so within the next year.

This year, it appears that they followed through. **The share of organizations already investing in deepfake defense rose 11 percentage points,**

**from 37% to 48%** from 2025 to 2026. And those new investors came from exactly where you'd expect. The share planning to invest within five months fell from 32% to 27%, within 6–11 months from 22% to 17%, and within 12–18 months from 7% to 5%. Based on the data, it appears last year's plans became this year's budgets.

## Q4. When does your organization plan to invest in deepfake protection, if ever?





# 48%

**of organizations are now actively investing in deepfake defense, up from 37% in 2025.**

Nonetheless, deepfake-related incidents themselves remain widespread. 84% of respondents reported at least one incident in the past 12 months, virtually unchanged from the 85% figure reported in 2025. However, fewer organizations are being hit again and again. **The share experiencing three or more incidents fell from 40% to 35%**, and the share hit more than five times fell from 12% to 9%.

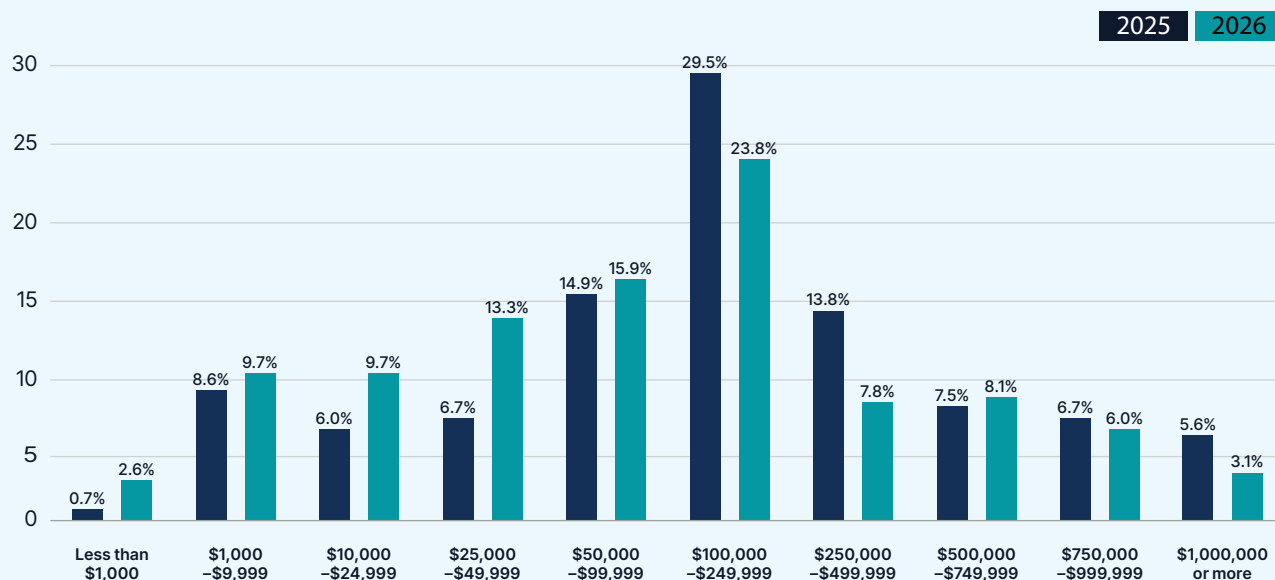
The bigger shift, though, is in what those incidents cost. Among organizations that reported losing money to deepfake or AI voice fraud, the loss distribution moved toward the lower end of the spectrum. **The share of affected organizations losing \$100,000 or more fell from 63% to 49%.** Losses of half a million dollars or more fell from 20% to 17%, and seven-figure losses were nearly cut in half, from 6% to 3%.

Taken together, the estimated average loss per affected organization dropped from roughly \$280,000 to roughly \$225,000, a decline of about 20%.<sup>3</sup>

<sup>3</sup> In 2025, loss amounts were collected only from respondents who reported a loss; in 2026, all respondents were asked, with a "\$0 / no known losses" option. Year-over-year comparisons are therefore limited to organizations reporting a nonzero loss, excluding "prefer not to say" and "unsure" responses. Averages are estimated from range midpoints.

## Q5. Please specify the estimated amount of money your organization has lost to deepfake or AI voice fraud in the past 12 months.<sup>3</sup>

\*Those whose organization lost money to a deepfake or AI voice fraud incident in the past 12 months



A pinch of salt: investment rose and losses fell over the same period, but survey data alone can't prove that one caused the other. Attacker tactics shift, reporting practices vary, and a single year is a short window. Still, the trend is encouraging, and it's what we'd hope to see when organizations put dedicated tools and processes in place against a specific threat. What hasn't eased, however, is concern.

# 96%

**of respondents say they're concerned about the threat deepfakes pose to their organizations (61% very concerned),** compared to 94% (63% very concerned) in 2025.

Deepfakes aren't going anywhere, and security and IT teams know it. But, as the rest of this report shows, deepfakes are no longer the only AI-driven threat competing for their attention. As things tend to go in the world of cybersecurity today, things appear to be going from bad to worse.



## SECTION 2

# AI Agents Have Arrived, and They Have the Keys

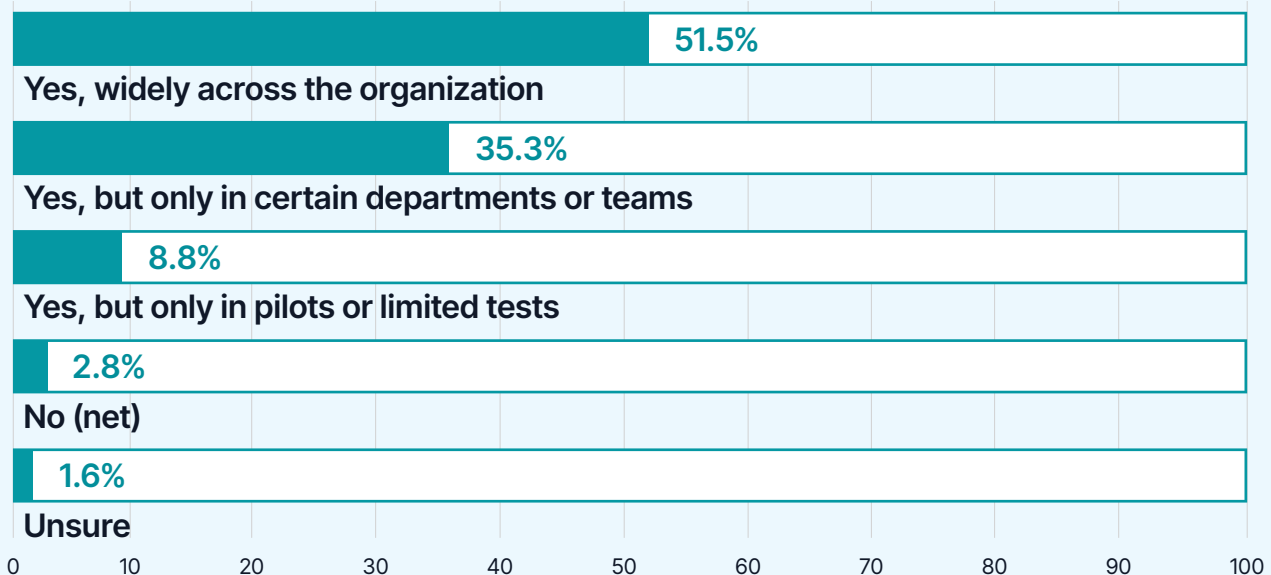
For the past few years, the conversation around AI in the enterprise has centered largely on AI assistants: tools that draft materials, summarize documents and meetings, and offer suggestions. These are often prompted and it is always a person making the final call. That's changing quickly. Today's AI agents and tools require far less human involvement and oversight, being able to act on employees' behalf: opening tickets, moving files, replying to messages, triggering workflows, and interacting with business systems. This year, we asked respondents how widely these tools have been adopted within their organizations. And the answer was a resounding "pretty much everywhere."

**96% of organizations now use AI agents or AI tools that can take actions on behalf of employees.** Only 3% say they don't. And these tools aren't confined to small experiments. Over half (52%) of respondents say use is widespread across the organization, another 35% say AI agents are in use in certain departments or teams, and just 9% are limited to pilots or limited tests.

# 96%

**of organizations use AI agents that can act on employees' behalf, and more than half use them widely.**

## Q6. Does your organization currently use AI agents or AI tools that can take actions on behalf of employees?

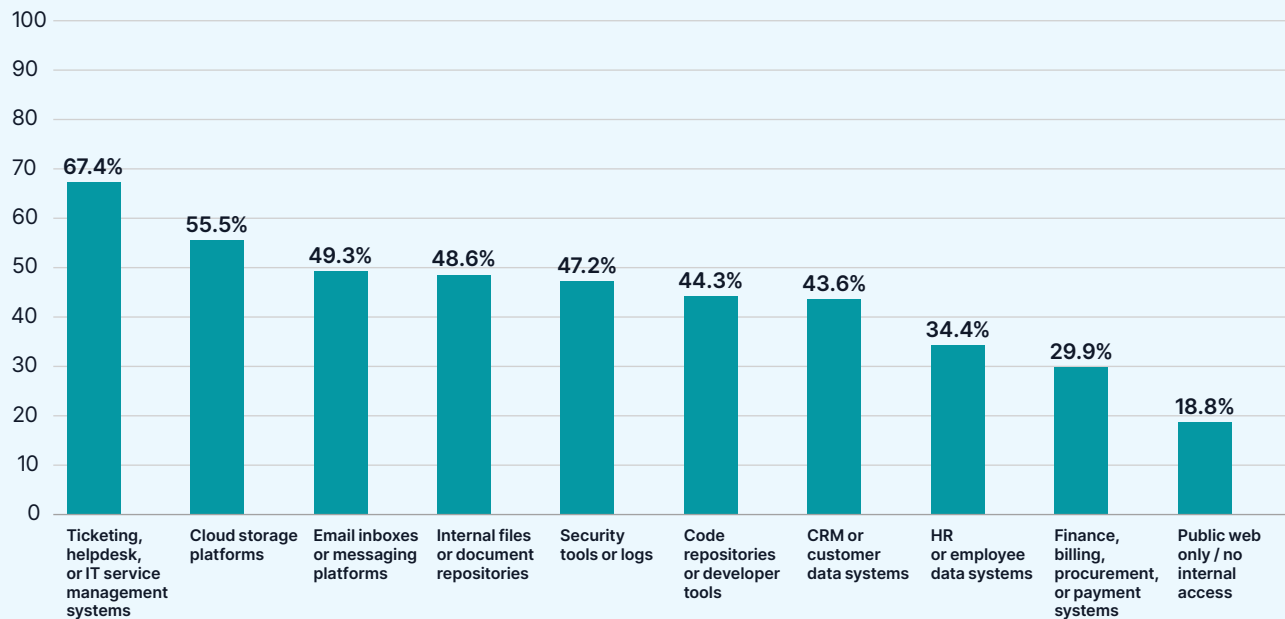


But adoption alone isn't the real risk. What makes these new tools dangerous is access and agency. An AI agent can only do as much damage as the systems it can reach and the actions they're enabled to take. And the systems these agents reach hold credentials, data, finances, and more. Among organizations using AI agents, the most common systems those agents can access are ticketing, helpdesk, or IT service management

systems (67%), cloud storage platforms (56%), email inboxes or messaging platforms (49%), internal files or document repositories (49%), and security tools or logs (47%). Code repositories (44%), CRM and customer data systems (44%), HR systems (34%), and even finance, billing, procurement, or payment systems (30%) follow close behind. Shockingly, **only 19% of respondents say their AI tools are limited to the public web with no internal access.**

## Q7. Which systems, if any, can AI agents or AI tools access in your organization today? (Select all that apply)

\*Those whose organization currently uses AI agents or AI tools that can take actions on behalf of employees



One finding among these figures deserves particular attention. Email and messaging platforms are the third most common environment AI agents can access, with nearly half of AI-using organizations (49%) granting that access. That means agents are reading inboxes, summarizing messages, and often acting on what they find.

For attackers, this opens up a whole new world of possibility. Phishing has always targeted the people who read email. Now it can target the software that reads email for them. A malicious message no longer has to fool an employee. If it can fool the agent sitting between that employee and the inbox, it has already succeeded.

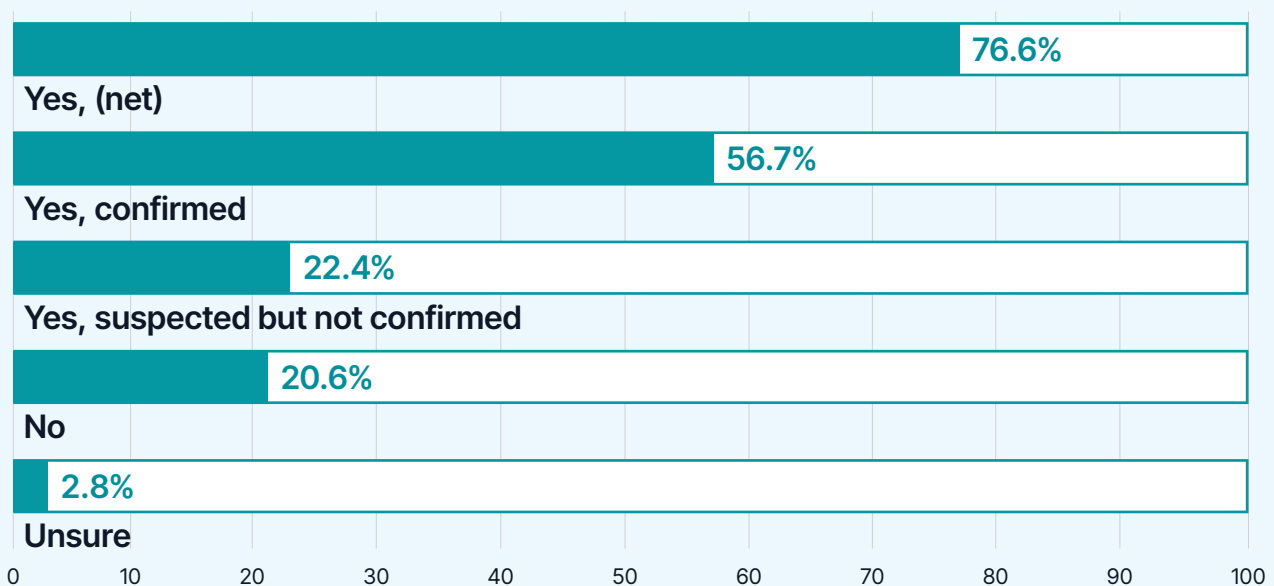
## SECTION 3

# Agents Are Getting Phished: 77% of Orgs Have Already Had an Incident

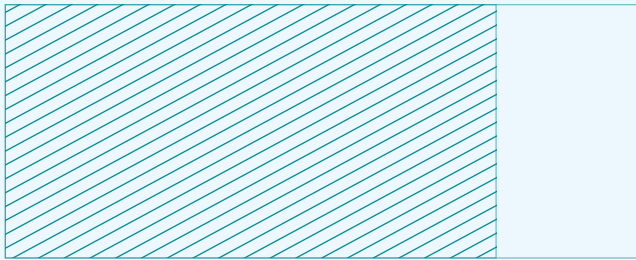
Unfortunately, AI agents don't just have the keys, attackers have also taken notice. **77% of respondents say that in the past 12 months, an AI agent or AI tool at their organization took an action based on instructions, prompts, emails, files, or content that later turned out to be malicious, misleading, or unauthorized.**

And these aren't just hunches. 57% of all organizations surveyed report at least one confirmed case, accounting for nearly three-quarters (74%) of organizations reporting an incident. An additional 20% suspect an incident was agent led but haven't confirmed it. **The exploitation of agentic AI is no longer a concern looming on the horizon. It is a real, present-day problem that is unquestionable and widespread.**

**Q9. In the past 12 months, has an AI agent or AI tool at your organization taken an action based on instructions, prompts, emails, files, or content that later turned out to be malicious, misleading, or unauthorized?**



Respondents could select more than one "yes" option, so sub-items sum to more than the net figure.



# 77%

**of organizations say an AI agent acted on malicious, misleading, or unauthorized content in the past year**

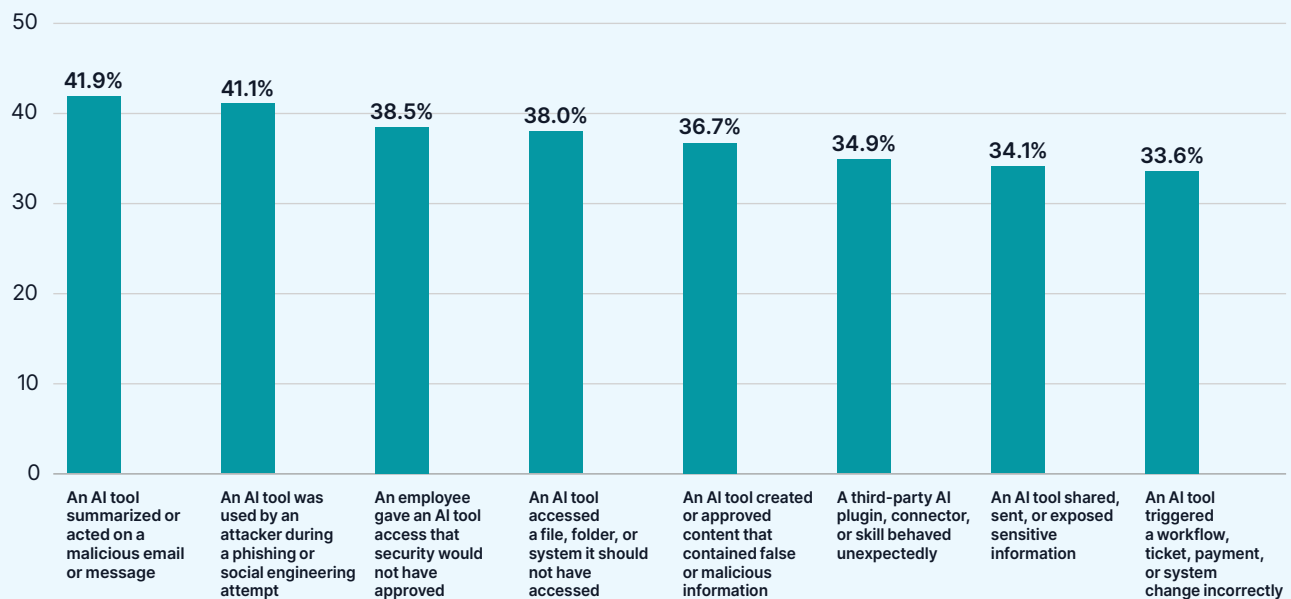
When we asked organizations that experienced an incident what happened, the answers were spread across every category we tested, with each cited by between 34% and 42% of respondents. No single failure mode dominates, a reminder of how many ways these tools can go wrong. But the two most common incident types both lead back to public enemy #1: email and social engineering:

- **An AI tool summarized or acted on a malicious email or message (42%)**
- **An attacker used an AI tool during a phishing or social engineering attempt (41%)**

Other notable incident types include an employee giving an AI tool access that security would not have approved (39%), an AI tool accessing a file, folder, or system it should not have (38%), and an AI tool triggering a workflow, ticket, payment, or system change incorrectly (34%).

## Q10. Which AI-agent-related incidents has your organization experienced or suspected in the past 12 months? (Select all that apply)

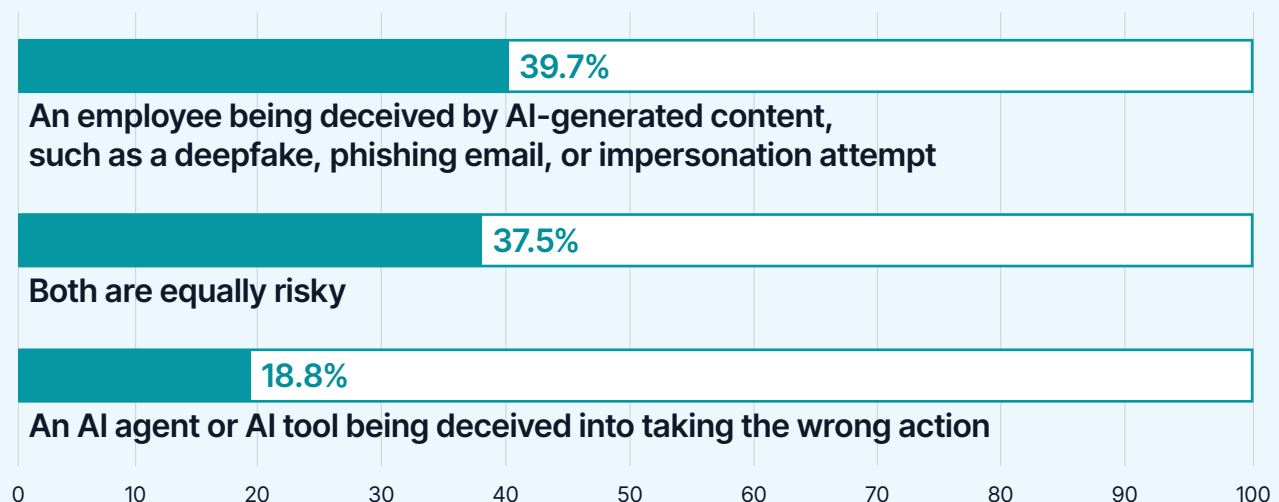
\*Those whose organization has had an AI agent or AI tool take an action based on content that later turned out to be malicious, misleading, or unauthorized in the past 12 months



Interestingly enough, professionals already know where the biggest agentic risk lies. Asked for the single most likely way an AI agent or tool could create a security incident at their organization, **respondents' top answer was "acting on a malicious email, prompt, file, or link" (23%)**, followed by accessing more systems or data than intended (17%). Meanwhile, a mere 2% said AI agents don't create meaningful security risks.

That has major implications for how organizations think about social engineering. When asked which poses the bigger risk over the next 12 months, 40% chose an employee being deceived by AI-generated content, 19% chose an AI agent being deceived into taking the wrong action, and 38% said both are equally risky. In other words, **56% see a deceived AI agent as at least as risky as a deceived employee**. That means social engineering now has two targets: people and AI.

**Q20. Out of the following, which is the bigger security risk to your organization over the next 12 months, if any?**



## SECTION 4

# The Ownership Gap: Security Is on the Sidelines

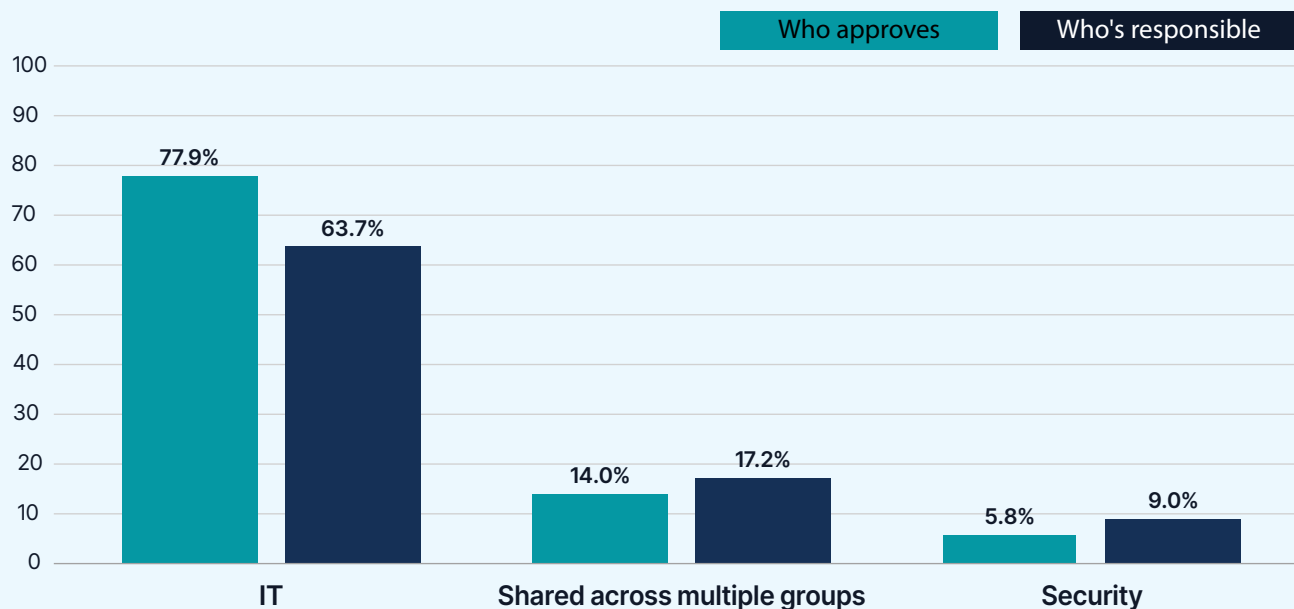
So why are AI agents being exploited at such a high rate? Survey data can't offer a definitive answer, but one finding stands out: **the teams responsible for security have surprisingly little say over which AI agents get deployed, what they can access, and who's accountable when things go wrong.**

Instead, it's IT that's clearly won the role of gatekeeper. Among organizations using AI agents, 78% say IT is primarily responsible for approving AI agents, AI assistants, plugins, and connectors before they're used. Approval is shared across multiple groups at 14% of organizations. Meanwhile, **security teams are the primary approver at just 6%.**

Rightfully so, IT also carries most of the accountability. If an AI agent were to cause a harmful security or business action, 64% of respondents say IT would bear the brunt of the responsibility, compared to 17% who say responsibility is shared and just 9% who say it falls to security. And when asked which department creates the greatest AI-agent-related risk, respondents pointed to IT yet again (66%), leaps and bounds ahead of any other group.

**Q8. Who, if anyone, is primarily responsible for approving AI agents, AI assistants, plugins, or connectors before they are used in your organization?**

**Q12. If an AI agent or AI tool caused a harmful security or business action, who, if anyone, would be primarily responsible at your organization?**



Q8 asked of organizations that currently use AI agents or AI tools (n=479); Q12 asked of all respondents (n=501).

**Security teams lead in the approval process of AI agents at just 6% of organizations.**

Despite the obvious objections, IT's leading role in approval is in some ways understandable. IT teams typically own the platforms, identity systems, and integrations these tools plug into, making it natural for them to be the ones to lead in adoption. But approving an AI agent is as much a risk decision for an enterprise as it is a functional one. An agent that can read email, browse cloud storage, or interact with payment systems is a new identity with real privileges, and decisions about what it can touch belong in front of the people whose job is to anticipate how it might be exploited or abused.

Which raises the question: is security's limited role in adopting these tools part of the problem? There's evidence to suggest it might be. **Among organizations that experienced an AI-agent-related incident, 39% say an employee gave an AI tool access that security would not have approved.** When security isn't part of the decision, access gets granted without a risk assessment or security review. And when those don't take place, it should come as little surprise when incidents occur down the line.

Still, there is good news here. Governance does exist, and its presence is nearly ubiquitous: none of the AI-using organizations we surveyed lack a formal approval process for AI agents, and only 2% of respondents say no one has decided who's responsible if something goes wrong. That may very well be because leadership is paying attention, too. **78% of respondents say their board or executive leadership requested a briefing on AI-agent or AI-tool security risks within the past year.** This is right on par with the 80% who reported board briefings on deepfake risk in 2025, suggesting a similar pattern of early correction could be in store for AI agents as well.

So, while the awareness and structure are in place, security's voice is conspicuously absent from decision making. With security professionals at the table (if not leading the adoption process), organizations would be better able to weigh the productivity gains of these tools against the security risks they pose.

---

## SECTION 5

# The Preparedness Paradox, Part Two

In last year's report, we identified what we called the preparedness paradox: despite 85% of organizations experiencing deepfake-related incidents, and a majority of them suffering financial losses, 99% of respondents claimed confidence in their defenses. This year, the same pattern has emerged around the topic of agentic AI.

On paper, organizations look well prepared for the agentic age:

- **93%** say their incident response plan covers incidents involving AI agents or AI tools (61% say it's fully covered).
- **93%** say their cybersecurity awareness training addresses safe use of AI tools or agents (55% in detail).
- **88%** test employee behavior around AI tools the same way they test phishing or social engineering risk (56% regularly).
- **84%** are confident they could identify whether an AI agent contributed to a security incident (53% very confident).

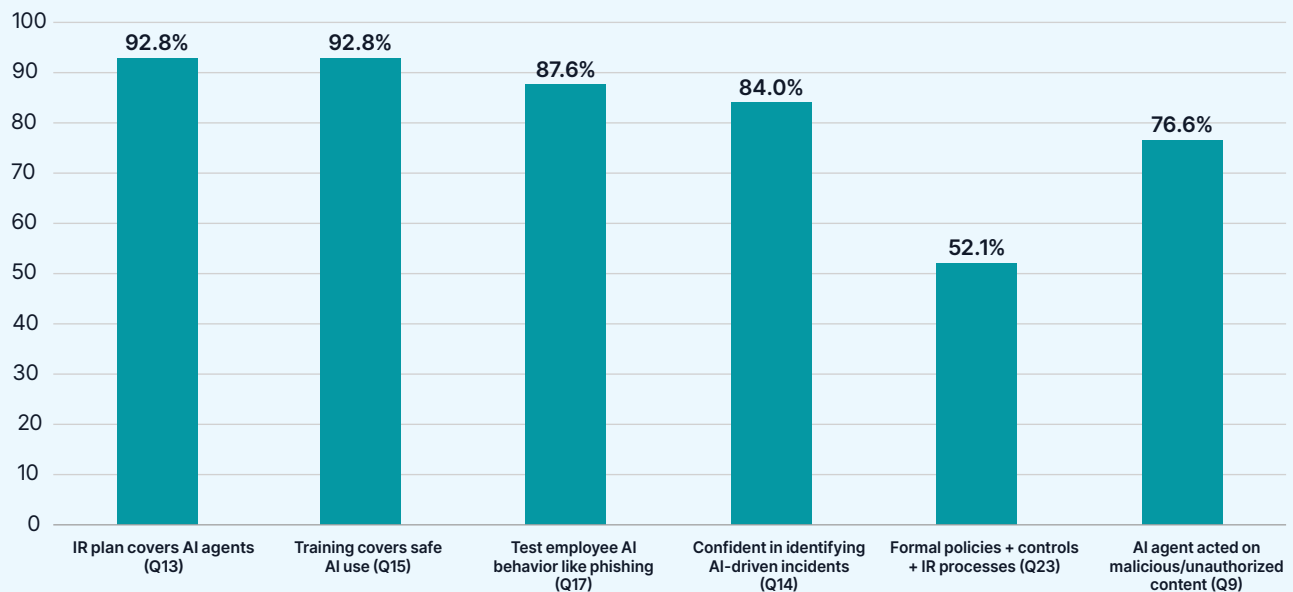
And yet, **77% have already had an AI agent act on malicious, misleading, or unauthorized content.** Only 52% of organizations describe

# 93%

**say their incident response plan covers AI agents, but 77% have already suffered from an AI-agent-related incident within the past 12 months.**

their approach to AI-agent security as having formal policies, controls, and incident response processes all in place. Another 35% have policies but limited technical controls. Policy and confidence are outpacing the technology and processes needed to back them up.

## "Readiness vs. Reality"



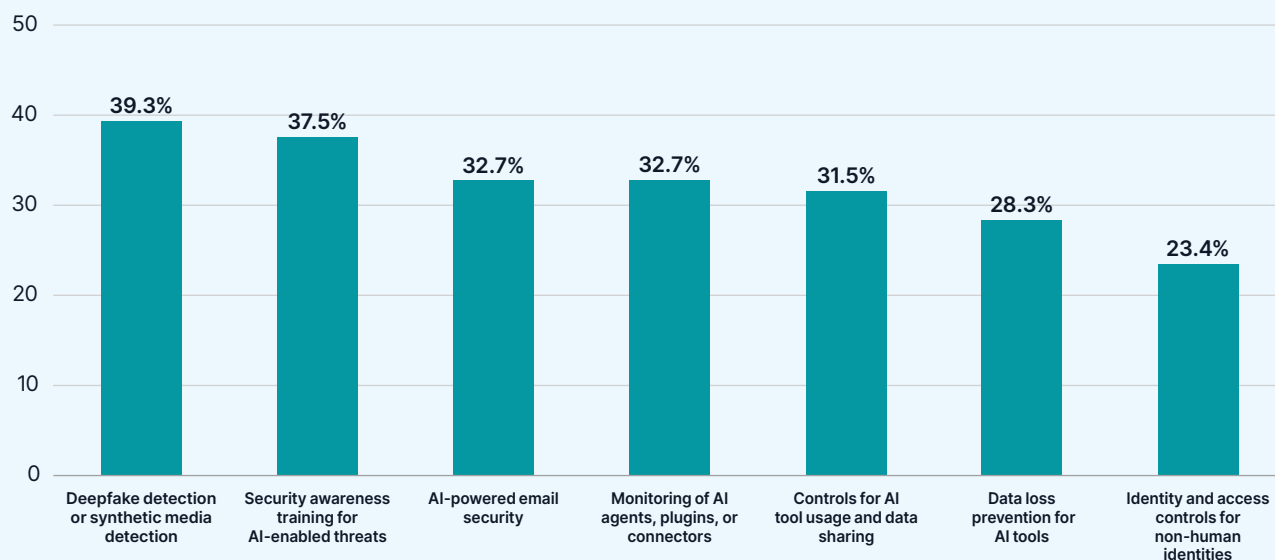
This is also where deepfakes re-enter the picture. Concern about deepfake-related cyberattacks is holding steady at 96%, and when asked which poses the bigger risk, an employee being deceived by AI-generated content, such as a deepfake, remains the single most-cited answer (40%). Deepfake detection also tops the list of planned security investments for the next 12 months (39%).

But the urgency is fading. **The share of respondents calling deepfake defense a “top priority” slipped 11 points year over year, from 71% to 60%.** And when asked which AI-related risk their organization is most likely to prioritize over the next year, deepfake-enabled phishing, BEC, and impersonation ranked third (16%), behind AI agents taking unauthorized actions (21%) and employees sharing sensitive data with AI tools (19%). That data exposes real potential

dangers: in a reactionary response, organizations may shift away from deepfakes and onto AI agents before their deepfake defenses have truly hardened. Last year’s progress seems to have come from sustained investment. Easing off now risks allowing that progress to unravel.

Investment intent, meanwhile, is broad but thin. When asked which controls they’re most likely to invest in over the next 12 months, respondents spread their answers widely, with the top five each landing between 31% and 39% and no clear consensus. Notably, AI-powered email security (33%) and monitoring of AI agents, plugins, or connectors (33%) are tied, and identity and access controls for non-human identities trail at 23%. That’s a modest showing for two controls aimed directly at the inbox, where most agent incidents begin, and at the access these agents already hold.

### Q19. Which controls, if any, is your organization most likely to invest in over the next 12 months? (Select up to 3)



## CONCLUSION

# Guidance for Today's Organizations

Last year's data delivered a clear and dire warning about deepfakes. This year's data shows what can happen when organizations take a warning seriously: investment followed, and losses became less severe. To stay ahead, organizations need to bring the same urgency to agentic AI, before losses from this new threat continue to mount; all while making sure not to turn a blind eye to deepfake-related threats. We at IRONSCALES recommend five fundamental steps moving forward:

## 1. Protect the inbox for agents as well as humans

Email is both the most common AI-agent incident type and the attack vector respondents fear most. As agents take on the work of reading, summarizing, and acting on messages, organizations need to secure the content agents consume, not just what employees click. AI-powered email security that analyzes intent, sender behavior, and context can stop malicious messages before a person or an agent ever acts on them.

## 2. Bring security into AI-agent approval

Make security a required approver for AI agents, plugins, and connectors, and clarify who is accountable when something goes wrong. When IT alone makes these decisions, what an organization stands to gain from new AI tools can take precedence over what it stands to lose. A shared approval process keeps productivity and risk in the same conversation.

### 3. Limit and monitor agent access

Apply the principle of least privilege to non-human identities, especially where agents touch finance, email, and security tooling. Grant agents only the access they need for the task at hand, monitor their actions regularly, and keep logs that allow for clear attribution if and when things go awry.

### 4. Test the paradox

Confidence is not the same as readiness. Routinely validate incident response plans and training with up-to-date, realistic simulations that include agent-targeted attacks, not only human-targeted ones. Make sure training and awareness programs reflect how work gets done in the agentic age, and stress-test the agents themselves through frequent red-teaming efforts.

### 5. Step up agentic AI security, without letting deepfake defenses lapse

Agentic AI is an immediate threat, and organizations should begin investing in the security to match, especially as it pertains to the inbox. Meanwhile, deepfakes aren't going anywhere. Keep investing in, and training for, deepfake threats while expanding those efforts to include AI agents. The progress organizations made against deepfakes this year is worth protecting.

## Not sure where to begin?

Reach out to the agentic security experts at [IRONSCALES](#) and discover how we can help you protect your people, and your AI agents, from AI-driven threats.

# Methodology

This original research was commissioned by IRONSCALES and conducted by [Censuswide](#). The survey polled a sample of 501 US-based IT + cybersecurity professionals (middle manager level and above). All participants were ages 25-55 and worked at companies of 1,000-10,000 employees. The data was collected between 09/09/2026 and 09/17/2026. Year-over-year comparisons are benchmarked against Censuswide's 2025 survey of 500 respondents with the same profile, collected between 09/08/2025 and 09/16/2025. Censuswide abides by and employs members of the Market Research Society and follows the MRS code of conduct and ESOMAR principles. Censuswide is also a member of the British Polling Council and a signatory of the Global Data Quality Pledge.

# About IRONSCALES

IRONSCALES is the leader in AI-powered email security, protecting over 18,000 global organizations from advanced email threats. As the pioneer of Adaptive AI, we catch the business email compromise (BEC), Account Takeover (ATO), and zero-day attacks that bypass other solutions, encrypt the sensitive data leaving the inbox, and automate the response. IRONSCALES is the first email security vendor verified under Anthropic's Cyber Verification Program, giving our platform frontier AI access to preempt attacks before they are sent. Combining AI and continuous human insights, we safeguard inboxes, unburden IT teams, and turn employees into a vital part of cyber defense across enterprises and managed service providers. IRONSCALES is headquartered in Atlanta, Georgia. To learn more, visit [www.ironcales.com](http://www.ironcales.com) or follow us on LinkedIn.

